Check for updates

Biometrical Journal **52** (2010) 61, zzz–zzz / DOI: 10.1002/bimj.200100000

Multiple imputation methods for handling incomplete longitudinal and clustered data where the target analysis is a linear mixed effects model

Md Hamidul Huque[*,1,2,3], **Margarita Moreno-Betancur** [1], **Matteo Quartagno** [4], **Julie A. Simpson** [5], **John B. Carlin** [1,2,5], and **Katherine J. Lee** [1,2]

[1] Murdoch Children's Research Institute, 50 Flemington Road, Parkville, VIC,3052, Australia.

[2] Department of Paediatrics, University of Melbourne, Parkville, VIC, 3052, Australia.

[3] University of New South Wales, Kensington, Kensington, NSW 2052, Australia.

[4] Institute for Clinical Trials and Methodology, University College London, 90 High Holborn, WC1V 6LJ.

[5] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, VIC, 3052, Australia.

Received 21 January 2019, revised zzz, accepted zzz

Multiple imputation (MI) is increasingly popular for handling multivariate missing data. Two general approaches are available in standard computer packages: MI based on the posterior distribution of incomplete variables under a multivariate (joint) model, and fully conditional specification (FCS), which imputes missing values using univariate conditional distributions for each incomplete variable given all the others, cycling iteratively through the univariate imputation models. In the context of longitudinal or clustered data, it is not clear whether these approaches result in consistent estimates of regression coefficient and variance component parameters when the analysis model of interest is a linear mixed effects model (LMM) that includes both random intercepts and slopes. In the current paper, we compared the performance of seven different MI methods for handling missing values in longitudinal and clustered data in the context of fitting LMMs with both random intercepts and slopes. We study the theoretical compatibility between specific imputation models fitted under each of these approaches and the LMM, and also conduct simulation studies in both the longitudinal and clustered data settings. Simulations were motivated by analyses of the association between body mass index (BMI) and quality of life (QoL) in the Longitudinal Study of Australian Children (LSAC). Our findings showed that the relative performance of MI methods vary according to whether the incomplete covariate has fixed or random effects and whether there is missingnesss in the outcome variable. We showed that compatible imputation and analysis models resulted in consistent estimation of both regression parameters and variance-components via simulation. We illustrate our findings with the analysis of LSAC data.

Key words: Fully conditional specification; Joint modelling; Missing data; Multiple imputation; Repeated measurement; clustered data

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1022/bimj.XXXXXXX

1 Introduction

Longitudinal and cluster-correlated data arise in many public health settings where data are collected from (i) individual participants repeatedly over time and (ii) from groups of individuals that are clustered within

[*] Corresponding author: e-mail: hamidul_b7@yahoo.com

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/bimj.201900051

This article is protected by copyright. All rights reserved.

Author Manuscript

Author Manuscript

natural units e.g, medical practices, geographical locations. Both of these settings have the common characteristic of correlated measurements either within an individual or within a cluster of individuals. Mixed-effects models are frequently used in the analysis of correlated data. However, the validity of the results obtained from such analyses may be compromised if some covariate values are missing (Laird, 1988).

Multiple imputation (MI) has become a popular tool for dealing with missing data in recent years (Rezvan et al., 2015). MI involves the generation of multiple copies of imputed datasets where missing values are replaced by imputed values sampled from their posterior predictive distribution (or an approximation to this) given the observed data. Each completed dataset is analyzed using the statistical model for the epidemiological question of interest, and the resulting estimates and standard errors are combined using Rubin's rules (Rubin, 1987). The theoretical basis of MI methods has been developed under the assumption that data are missing at random (MAR), which requires that the probability of data being missing does not depend on the unobserved data, conditional on the observed data (Sterne et al., 2009). If the data are MAR, a correctly implemented MI method can produce unbiased and asymptotically efficient estimates of regression parameters and their standard errors. Correct implementation requires compatibility between the imputation and analysis models. Formally, a set of conditional models are called compatible if there exists a joint density function that generates them (Meng, 1994).

Two general approaches for implementing MI in the presence of multiple incomplete variables are available in the literature: MI based on the joint posterior distribution of incomplete variables, often referred to as joint modeling (JM) (Schafer, 1997), and fully conditional specification (FCS; also known as sequential regression and MI using chained equation (MICE)) (Raghunathan et al., 2001; Van Buuren et al., 2006). The JM approach assumes that the incomplete variables follow a multivariate distribution, usually a multivariate normal distribution in which case the method is referred to as multivariate normal imputation (Schafer, 1997). FCS, on the other hand, imputes missing values using univariate conditional distributions for each incomplete variable given all the other variables in the imputation model, cycling iteratively through the univariate imputation models (Raghunathan et al., 2001; Van Buuren et al., 2006). Both the JM and FCS approaches were originally proposed for imputing missing values in cross-sectional settings with independent observations, and subsequently various extensions have been proposed in the literature to accommodate longitudinal and correlated data.

MI methods developed to impute missing values in both the cluster-correlated and longitudinal data settings include a joint multivariate linear mixed effects model (LMM) approach (JM-MLMM) (Schafer and Yucel, 2002), implemented in the *pan* software in R. There is also an FCS adaptation of Schafer and Yucel's approach (FCS-LMM) implemented in the *mice.impute.2lpan* function of the mice package in R (Van Buuren and Groothuis-Oudshoorn, 2011). Both the JM-MLMM and FCS-LMM approaches assume a constant residual variance across all clusters. Subsequently, Yucel and Van Buuren et al. extended the JM-MLMM and FCS-LMM approaches to allow for heteroscedastic (cluster-specific) random covariance matrices and residual error variances, respectively (Yucel, 2011; Van Buuren et al., 2011), hereby denoted as JM-MLMM-het and FCS-LMM-het. Both JM-MLMM and JM-MLMM-het, and their FCS adaptations (FCS-LMM and FCS-LMM-het), assume normal distributions for the incomplete variables. In practice, incomplete variables may be a mixture of continuous and categorical variables, so the assumption of normality may not be realistic. Goldstein et al. proposed an extension of the JM-MLMM approach which uses latent normal (LN) variables to impute a mixture of discrete, normal and non-normal continuous variables, referred to herein as JM-MLMM-LN (Goldstein et al., 2009). Asparouhov and Muthén suggested a method similar to JM-MLMM-LN where all variables in the imputation models are treated as outcomes, regardless of missing data pattern, hereby denoted as full joint (JM-FJ) model (Asparouhov and Muthén, 2010). More recently, Goldstein et al. proposed a further extension of JM-MLMM-LN where the imputation model is defined as the product of the substantive model and the joint distribution of the covariates, to ensure congeniality (substantive model compatible, denoted JM-SMC) (Goldstein et al., 2014) . The JM-MLMM-LN and JM-SMC approaches have been implemented in the REALCOM and Stat-JR software packages, respectively (see *http://www.bristol.ac.uk/cmm/software/*) and both were later adopted in the R software package *jomo* (Quartagno and Carpenter, 2016). The *jomo* implementations for JM-MLMM-LN

This article is protected by copyright. All rights reserved.

Author Manuscript

and JM-SMC allow a random covariance matrix and hence are denoted as JM-MLMM-LN-het and JM-SMC-het. Similar efforts have been made to extend both the FCS-LMM and FCS-LMM-het methods to impute categorical data using either generalized LMM (GLMM)-based MI methods (Resche-Rigon and White, 2016; Zhao and Yucel, 2009) or LN variables (FCS-LMM-LN and FCS-LMM-LN-het) (Enders et al., 2017).

In the special case of longitudinal data collected at equal intervals, standard cross-sectional implementations of MVNI and FCS can be employed to impute missing values by treating the time-dependent longitudinal measurements as distinct variables (Schafer, 1997; Van Buuren et al., 2006); we denote these as JM-MVN and FCS-Standard, respectively. These single-level MI methods can also be used for cluster-correlated data by including cluster-specific indicator variables to capture the within-cluster correlation – known as 'fixed cluster imputation' (Reiter et al., 2006).

Although similar MI methods can be used to impute missing values in both longitudinal and clustered data settings, the performance of these methods may differ according to the intra-subject/intra-cluster association between outcome and incomplete variables in the analysis model particularly in the situation when both the outcome and covariates associated with random effects contain missing values. In the longitudinal setting, random slopes (i.e., random coefficients for covariates) are usually associated with the time variable only, which is generally fully observed, but covariates with random slopes in the context of clustered data may be incomplete. Furthermore, it is unclear how important these differences are in practice as currently available comparisons of the various MI methods in the literature are limited to either clustered or longitudinal data settings with little theoretical consideration.

In the context of cluster-correlated data, Grund et al. compared two different modeling strategies with JM-MLMM: (i) a multivariate LMM with a so-called reverse random coefficients model assuming that the outcome is fully observed (this model regresses covariates on the outcome with the outcome having random effects if the covariate has them in the analysis model) for imputing missing data in covariates and (ii) a multivariate LMM with random intercepts only (thus ignoring random slopes in the outcome model) for imputing missing data in both covariates and outcome (Grund et al., 2016). They noted that the reverse random coefficients model provided unbiased estimates of the regression and variance components, but the second model performed poorly for the estimation of the random slope variance. Similar findings were also observed by Enders and colleagues (Enders et al., 2016) who compared JM-MLMM, FCS-LMM-het and fixed cluster imputation when both the outcome and covariates contain missing data. They reported that the FCS-LMM-het approach exhibited better performance than the other methods especially when both the outcome and covariate were incomplete in a random intercept and slope analysis model. Audigier and colleagues (Audigier et al., 2018) recently compared a number of methods including fixed cluster imputation, JM-MLMM, FCS-LMM-het, and JM-MLMM-LN-het in the context of cluster-correlated data and reported that all of these methods provided reliable estimation of the regression parameters but JM-MLMM and fixed cluster imputation approaches severely under-estimated the variance components. No such comparison in the context of longitudinal data, where the analysis model of interest is a random intercept and time-slope model, is available in the literature. Recently, we compared 12 different MI approaches for imputation of incomplete longitudinal data where the analysis model of interest is a LMM with subject-specific random intercept only (Huque et al., 2018). We showed that both standard MI methods (JM-MVN and FCS-Standard) and LMM-based approaches (JM-MLMM, JM-MLMM-LN, FCS-LMM and FCS-LMM-LN), provided consistent estimates of the regression and variance component parameters. However, these results may not be generalizable to a random intercept and slope analysis model. Moreover, all the above comparisons are empirical and no theoretical justification for the observed sub-optimal results is available.

The motivation for this study was an analysis of the Longitudinal Study of Australian Children (LSAC) that explored (a) the association between body mass index (BMI) and health related quality of life (QoL) for children over time and (b) whether the association between early BMI and QoL in later life varied across geographical location. Attrition and non-response make these data a natural candidate for analysis using MI, but no clear guideline was available on the selection of the appropriate MI method. In the current

This article is protected by copyright. All rights reserved.

Author Manuscript

paper, we study the properties of available MI methods, both theoretically and via simulations based on these examples, and we also perform an analysis of the LSAC data. As both BMI and QoL are continuous measures, we restrict our comparisons to the approaches where all variables in the MI model are continuous. This simplifies the study of theoretical compatibility between specific imputation models fitted under each of these MI approaches and the analysis model and reduces the number of competitive MI methods, as under this restriction the MI methods with latent normal variables (JM-MLMM-LN, FCS-LMM-LN and FCS-LMM-LN-het) are identical with those that treat all the variables as continuous (JM-MLMM, FCS-LMM, and FCS-LMM-het, respectively). Our study of compatibility confirms that MI approaches result in consistent estimates of regression parameters when the imputation model is compatible with the analysis model. The results from the LSAC data analysis are also in agreement with those seen in the simulation study.

The structure of the article is as follows: Section 2 describes LSAC and the analysis models of interest. Sections 3 and 4 present a theoretical exploration of the compatibility of different MI methods and a linear mixed model with random intercept and slopes as analysis model in the context of longitudinal and cluster-correlated data, respectively. Section 5 describes and presents the results of our simulation study. The application to the LSAC data is presented in Section 6. We conclude with a general discussion in Section 7. The Web Appendices give detailed proofs, as needed.

2 Methods

2.1 Analysis models of interest

Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^{\mathrm{T}}$ be the n_i-repeated measures of a continuous outcome for individual $i \in (1, 2, ..., n)$, and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{in_i})^{\mathrm{T}}$ and $\boldsymbol{t}_i = (t_{i1}, t_{i2}, ..., t_{in_i})^{\mathrm{T}}$ represent repeated measures of a continuous covariate and the measurement times, respectively. Suppose the association between the repeated measured outcome and covariates can be expressed using the following LMM

$$\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{t}_i = \beta_0 + \beta_1\boldsymbol{x}_i + \beta_2\boldsymbol{t}_i + \boldsymbol{b}_{0i} + \boldsymbol{b}_{1i}\boldsymbol{t}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, 2, ...n \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ is the vector of fixed-effects, $\boldsymbol{b}_i = (\boldsymbol{b}_{0i}, \boldsymbol{b}_{1i}) \sim N(\boldsymbol{0}, \boldsymbol{G})$ denotes the random effects vector and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{in_i}) \sim N(\boldsymbol{0}, \Phi_i = \sigma^2_{\varepsilon_i}\boldsymbol{I})$, where $\boldsymbol{I}$ is the $n_i \times n_i$ identity matrix. The LMM in (1) typically assumes that the residual error, $\boldsymbol{\varepsilon}_i$ and random effects $\boldsymbol{b}_i$ are independent of each other. Thus the marginal distribution of $\boldsymbol{y}_i$ is $MVN(\mu_{\boldsymbol{y}_i} = \beta_0 + \beta_1\boldsymbol{x}_i + \beta_2\boldsymbol{t}_i, \boldsymbol{\Sigma}_{y_i} = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^{\mathrm{T}} + \Phi_i)$, where $Z_i = (\boldsymbol{1}, \boldsymbol{t}_i)^{\mathrm{T}}$ is a $n_i \times 2$ matrix with the first column having all elements equal to 1. This LMM models the longitudinal trajectory for each subject over time.

A similar model can also be applied to clustered data where the effect of some covariates on the outcome are allowed to vary from cluster to cluster. In the clustered data setting, the LMM with a random intercept and slope might take the following form

$$\boldsymbol{y}_i|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i} = \alpha_0 + \alpha_1\boldsymbol{x}_{1i} + \alpha_2\boldsymbol{x}_{2i} + \boldsymbol{a}_{0i} + \boldsymbol{a}_{2i}\boldsymbol{x}_{2i} + \xi_i, \qquad i = 1, 2, ...m; \tag{2}$$

where $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ are vectors of measurements of covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively within cluster $i \in (1, 2, ...m)$, assumed to be associated with the outcome, $\boldsymbol{y}_i$.

The estimation of parameters for the above LMMs can be carried out in similar fashion if all the variables in the model are complete. However, in the presence of incomplete data in the covariates the above two classes of models could differ: in the longitudinal setting, the covariate associated with the random slope, the measurement times $\boldsymbol{t}$ is generally observed, while in the clustered data settings, the covariate ($\boldsymbol{x}_2$) associated with a random slope may be incomplete. To assess the performance of the MI approaches in these distinct situations where an LMM with random intercepts and slopes is the analysis model of interest, we evaluated their performance under the following four scenarios: in the case of a longitudinal study where (i) only the covariate $\boldsymbol{x}$ is incomplete and (ii) both the covariate $\boldsymbol{x}$ and outcome $\boldsymbol{y}$ are incomplete;

This article is protected by copyright. All rights reserved.

Author Manuscript

and in the context of clustered data where (iii) only the covariate $\boldsymbol{x}_2$ is incomplete and (iv) both covariate $\boldsymbol{x}_2$ and outcome $\boldsymbol{y}$ are incomplete.

In the next two sections we study the theoretical properties of various MI methods available for imputing longitudinal and clustered data, in particular, we examine the potential for compatibility of each imputation model with the analysis model of interest.

3 MI methods for missing data in longitudinal settings

In longitudinal studies, data from the same individuals are collected repeatedly over time. Longitudinal data can be arranged in the wide format (new variable for each repeated measurement) if measurements occur at the same time-points for all individuals (i.e., the dataset is balanced) or in the long format (where repeated measurements are stacked). The wide format data can be imputed using standard cross-sectional imputation models (JM-MVN and FCS-Standard) by assuming the repeated assessments of the same variable are distinct variables, while imputation with the long format data requires use of multilevel imputation models.

3.1 JM-MVN

JM-MVN can be applied if we have balanced longitudinal data by treating all the repeated measurements of time-dependent variables as distinct. This method assumes a multivariate normal distribution for all of the incomplete variables. More specifically, assume that both the time-dependent covariates and outcome for individual $i \in (1, 2, ..., n)$ measured on T occasions, where $t = (1, 2, ...T)$ represents the vector of time-points when the measurements took place. If both covariate $\boldsymbol{x}$ and outcome $\boldsymbol{y}$ are incomplete, then JM-MVN assumes that $(\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_T, \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and an unstructured variance-covariance matrix, respectively.

We study the congeniality between JM-MVN and the analysis model (1) in the setting where the covariate $\boldsymbol{x}_i$ also follows a LMM defined as

$$\boldsymbol{x}_i|\boldsymbol{t}_i \quad = \quad \gamma_0 + \gamma_1 \boldsymbol{t}_i + \boldsymbol{u}_{0i} + \boldsymbol{u}_{1i}\boldsymbol{t}_i + \boldsymbol{\epsilon}_i, \tag{3}$$

where $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \Upsilon)$ and $\boldsymbol{u}_i = (\boldsymbol{u}_{0i}, \boldsymbol{u}_{1i}) \sim N(\boldsymbol{0}, \boldsymbol{D})$, where Υ and $\boldsymbol{D}$ are the covariance matrices currently left unspecified. As both the conditional distributions of $(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{t}_i)$ and $(\boldsymbol{x}_i|\boldsymbol{t}_i)$ are Gaussian, the joint distribution of $(\boldsymbol{y}_i, \boldsymbol{x}_i|\boldsymbol{t}_i)$ is also Gaussian. Since we are assuming that the data are collected for an equal number of visits at fixed time intervals for all individuals, the joint distribution of $(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{t})^{\mathrm{T}} = (y_1, y_2, ...y_T, x_1, x_2, ...x_T|1, 2, ..T)$ is normal and given by

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{pmatrix} |\boldsymbol{t} \Bigg) \quad = \quad N\left(\boldsymbol{\mu} = \begin{pmatrix} \beta_0 + \beta_1(\gamma_0 + \gamma_1 \boldsymbol{t}) + \beta_2 \boldsymbol{t} \\ \gamma_0 + \gamma_1 \boldsymbol{t} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \beta_1 \Sigma_x \beta_1^{\mathrm{T}} + \Sigma_y & \beta_1 \Sigma_x \\ \Sigma_x \beta_1 & \Sigma_x \end{pmatrix}\right) \tag{4}$$

[see the Appendix A.1 for proof]. Therefore, the joint distribution assumed by JM-MVN in scenario (ii) is compatible with the joint distribution implied by analysis model (1).

Scenario (i), where there is missing data only for $\boldsymbol{x}$, is a special case of scenario (ii), hence JM-MVN will be compatible with analysis model (1) for this scenario too.

3.2 JM-MLMM

Instead of treating repeated measurements as distinct variables, Schafer and Yucel suggested using a multivariate LMM for imputing several incomplete longitudinal variables (Schafer and Yucel, 2002). Under scenario (ii) this method imputes missing data from the following multivariate LMM:

$$\left(\begin{matrix} \boldsymbol{x}_i \\ \boldsymbol{y}_i \end{matrix} \; |\boldsymbol{t}_i \right) = \begin{pmatrix} \beta_{0(x)} + \beta_{1(x)}\boldsymbol{t}_i + \boldsymbol{b}_{0(x)i} + \boldsymbol{b}_{1(x)i}\boldsymbol{t}_i + \boldsymbol{\varepsilon}_{(x)i} \\ \beta_{0(y)} + \beta_{1(y)}\boldsymbol{t}_i + \boldsymbol{b}_{0(y)i} + \boldsymbol{b}_{1(y)i}\boldsymbol{t}_i + \boldsymbol{\varepsilon}_{(y)i} \end{pmatrix} \tag{5}$$

This article is protected by copyright. All rights reserved.

Author Manuscript

where $\begin{pmatrix} \boldsymbol{b}_{0(x)i} & \boldsymbol{b}_{1(x)i} \\ \boldsymbol{b}_{0(y)i} & \boldsymbol{b}_{1(y)i} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Psi})$ and $\begin{pmatrix} \boldsymbol{\varepsilon}_{(x)i} \\ \boldsymbol{\varepsilon}_{(y)i} \end{pmatrix} \sim N[\mathbf{0}, (\Sigma \otimes \boldsymbol{I})]$. The covariance matrix $\boldsymbol{\Psi}$ has dimension 4×4 and the Kronecker product notation indicates that the $\boldsymbol{\varepsilon}_{(x)i}$ and $\boldsymbol{\varepsilon}_{(y)i}$ are independently distributed as $N(0, \Sigma)$. With some algebra we can show that analysis model (1) can be obtained as a special case of the conditional model for the outcome given the covariate, $\boldsymbol{x}$, under the bivariate joint distribution defined in (5) [see Appendix A.2 for proof]. Hence the JM-MLMM model would be compatible with the analysis model of interest under scenario (ii).

Under scenario (i) i.e., when only covariate $\boldsymbol{x}$ contains missing data, the imputation model under JM-MLMM is given by

$$\boldsymbol{x}_i | \boldsymbol{y}_i, \boldsymbol{t}_i = \beta_{0(x)} + \beta_{1(x)} \boldsymbol{y}_i + \beta_{2(x)} \boldsymbol{t}_i + \boldsymbol{b}_{0(x)i} + \boldsymbol{b}_{1(x)i} \boldsymbol{t}_i + \boldsymbol{\varepsilon}_{(x)i} \tag{6}$$

where $\boldsymbol{\varepsilon}_{(x)i} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{(x)})$ and $\boldsymbol{b}_{(x)i} = (\boldsymbol{b}_{0(x)i}, \boldsymbol{b}_{1(x)i}) \sim N(\mathbf{0}, \boldsymbol{\Psi}_{(x)})$ with $\boldsymbol{\Sigma}_{x|y} = \boldsymbol{Z}_i \boldsymbol{\Psi}_{(x)} \boldsymbol{Z}_i^{\mathrm{T}} + \Sigma$. Thus, under scenario (i), JM-MLMM would be compatible with the substantive model if both of the conditional models $\boldsymbol{x}_i | \boldsymbol{y}_i, \boldsymbol{t}_i$ and $\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{t}_i$ lie in the subspace determined by the joint model $\begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{y}_i \end{pmatrix} | \boldsymbol{t}_i \Big)$. It can be shown that imputation model (6) is compatible with analysis model (1) if $\beta_{1(x)}^{\mathrm{T}} \boldsymbol{\Sigma}_{x|y}^{-1} = \beta_1 \boldsymbol{\Sigma}_y^{-1}$ [see the Appendix A.2 for proof]. Similar conditions for two linear regressions to be compatible when the target joint distribution is bivariate normal have also been noted (Zhu and Raghunathan, 2015) and (Liu et al., 2014). The current paper extends those results to the context of LMM.

3.3 JM-FJ

Asparouhov and Muthén suggested an alternative to the JM-MLMM-LN (Goldstein et al., 2009) where the data are imputed using an unrestricted model, where all variables in the imputation models are treated as outcome, regardless of missing data pattern, hereby denoted as full joint (JM-FJ) model (Asparouhov and Muthén, 2010) . The JM-FJ method under both scenario (i) and (ii) is given by

$$\begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{x}_i \\ \boldsymbol{t}_i \end{pmatrix} = \begin{pmatrix} \beta_{0(y)} + \boldsymbol{b}_{(y)0i} + \boldsymbol{\varepsilon}_{(y)i} \\ \beta_{0(x)} + \boldsymbol{b}_{(x)0i} + \boldsymbol{\varepsilon}_{(x)i} \\ \beta_{0(t)} + \boldsymbol{b}_{(t)0i} + \boldsymbol{\varepsilon}_{(t)i} \end{pmatrix}, \tag{7}$$

where $\begin{pmatrix} \boldsymbol{b}_{(y)0i} \\ \boldsymbol{b}_{(x)0i} \\ \boldsymbol{b}_{(t)0i} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Omega}_u)$ and $\begin{pmatrix} \boldsymbol{\varepsilon}_{(y)i} \\ \boldsymbol{\varepsilon}_{(x)i} \\ \boldsymbol{\varepsilon}_{(t)i} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Omega}_\varepsilon)$. This model imposes the same random-effect structure for all variables, decomposing the variance into within and between-individual components. In longitudinal studies, data are often collected at fixed time intervals for all individuals, and therefore, it may not be sensible to assume between-individual variability for the time variable (or corresponding latent variable). The JM-FJ approach has a large number of parameters and convergence is often difficult to achieve. Moreover, it can be shown that the joint distribution implied by JM-FJ (7) is not compatible with the substantive model (1) [see Appendix A.3 for proof]. This uncongeniality is due to the fact that the JM-FJ does not accommodate the variability in the slope across individuals. Because of this non-congeniality, in our simulation studies we also examine whether assuming heteroscedastic covariance matrices in the imputation model may improve the estimation of the variance components by allowing for subject-specific correlations (JM-FJ-het).

3.4 JM-SMC

Goldstein, Carpenter and Browne (2014) extended JM-MLMM-LN to handle missing data in both covariates and outcomes in multilevel models while ensuring that the imputation model is compatible with the substantive model (Goldstein et al., 2014). We refer to this as the substantive-model-compatible joint modeling approach (JM-SMC). In this formulation, the joint imputation model is defined as a product of the

This article is protected by copyright. All rights reserved.

joint distribution of covariates and the analysis model (i.e., conditional model for the outcome given the covariates). Specifically, the JM-SMC approach defines the joint distribution of $\begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{y}_i \end{pmatrix}|\boldsymbol{t}_i$ as

$$\begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{y}_i \end{pmatrix}|\boldsymbol{t}_i = (\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{t}_i) \times (\boldsymbol{x}_i|\boldsymbol{t}_i)\,, \tag{8}$$

where $(\boldsymbol{x}_i|\boldsymbol{t}_i = \beta_{0(x)} + \beta_{(x)}\boldsymbol{t}_i + \boldsymbol{b}_{0(x)i} + \boldsymbol{b}_{1(x)i}\boldsymbol{t}_i + \boldsymbol{\varepsilon}_{(x)i})$ with $\boldsymbol{b}_{(x)} \sim N(\mathbf{0}, \theta_u)$ and $\boldsymbol{\varepsilon}_{(x)i} \sim N(\mathbf{0}, \Theta_{\boldsymbol{\varepsilon}})$. The JM-SMC thus ensures compatibility under both scenarios (i) and (ii). Similarly to JM-FJ, in our simulation we also assume heteroscedastic covariance matices for the imputation using JM-SMC, and we labeled this JM-SMC-het.

3.5 FCS-Standard

Similarly to JM-MVN, FCS-Standard can be applied only in the setting with regular measurement timepoints, by treating all the repeated measurements of time-dependent variables as distinct variables. Specifically, this approach involves a conditional imputation model for each time-and-variable-specific measurement given the remaining measurements and variables. When considering only continuous outcome and covariates, as in this manuscript, FCS-Standard is implemented using linear regression models without interactions between covariates for the univariate imputation models. In this situation, FCS-Standard and JM-MVN are equivalent (see proposition 1 of (Hughes et al., 2014)). Given we have shown that JM-MVN is compatible with analysis model (1) under model (3) for the incomplete covariate, FCS-Standard will also be compatible with the analysis model (1) in both scenarios under these conditions.

3.6 FCS-LMM

Instead of treating repeated measurements as distinct variables, the FCS-LMM method uses a LMM for imputing missing values in each incomplete time-dependent variable given all the others, cycling iteratively through the univariate imputation models. Specifically, the Gibbs sampler cycles through the univariate LMMs assuming homogeneous within-subject variance, which is a special case of a multivariate LMM (5). That is, it uses the same imputation models as JM-MLMM with only one variable considered incomplete at a given iteration. Under scenarios (i) and (ii), this method will be compatible with the analysis model if the compatibility condition (derived in 3.2) is satisfied.

3.7 FCS-LMM-het

Similarly to FCS-LMM, FCS-LMM-het imputes each time dependent incomplete variable using a LMM. However, this method allows a subject-specific residual error variance. Under this approach, the imputation model for covariate $\boldsymbol{x}$ associated with the $i'th$ subject of interest is given by

$$(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{t}_i, \boldsymbol{b}_i) = N\left(\beta_{0i(x)} + \beta_{1i(x)}\boldsymbol{y}_i + \beta_{2i(x)}\boldsymbol{t}_i, \Sigma_{ix|y} = \sigma^2_{ix}\boldsymbol{I}_{n_i}\right), \tag{9}$$

where $\beta_{0i(x)} = \beta_{0(x)} + \boldsymbol{b}_{0(x)i}, \beta_{1i(x)} = \beta_{1(x)} + \boldsymbol{b}_{1(x)i}$ and $\beta_{2i(x)} = \beta_{2(x)} + \boldsymbol{b}_{2(x)i}$. Note the FCS-LMM-het approach assumes random slopes for each variable in the imputation model. Analysis model (1) can be re-written as

$$(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{t}_i, \boldsymbol{b}_i) = N\left((\beta_0 + \boldsymbol{b}_0 i) + \beta_1\boldsymbol{x}_i + (\beta_2 + \boldsymbol{b}_{1i})\boldsymbol{t}_i, \Sigma_{y|x} = \sigma^2_y\boldsymbol{I}_{n_i}\right)$$

Using similar arguments as with FCS-LMM, it can be shown that under scenario (i) FCS-LMM-het would be compatible with the analysis model if both the conditional model $\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{t}_i$ and $\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{t}_i$ lie in the subspace determined by the joint model $\begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{y}_i \end{pmatrix}|\boldsymbol{t}_i$. It can thus be shown that the imputation model (9) is compatible with the analysis model (1) if $\beta^{\mathrm{T}}_{1i(x)}\boldsymbol{\Sigma}^{-1}_{ix|y} = \beta_1\boldsymbol{\Sigma}^{-1}_y$, which is very similar to the condition

Author Manuscript

This article is protected by copyright. All rights reserved.

derived in (3.2). Hence, this method will be compatible with the analysis model (1) under both scenarios (i) and (ii), if the above compatibility condition is satisfied.

In summary, for longitudinal studies, we anticipate that all of the above methods, except the JM-FJ, will provide consistent estimates of regression and variance components.

4 MI models for missing data in cluster-correlated data.

In the cluster-correlated settings data are arranged in a long format by stacking data from each cluster. The MI methods that can be carried out are i) standard JM and FCS approaches using a total of m-1 indicator variables representing allocation of m clusters as a fixed factor in the model (fixed cluster imputation) (Reiter et al., 2006; Enders et al., 2016), or ii) a multilevel imputation method.

The use of indicator variables in fixed cluster imputation preserves the difference in intercept between clusters. However, an interaction between the indicator variables and the incomplete variables will also be needed to accommodate the random slope variation if the covariate(s) associated with random slopes are incomplete. However, such analysis requires estimation of a large number of parameters and hence is computationally demanding and often infeasible particularly with large number of clusters of small sizes (Enders et al., 2016). In contrast, the multilevel imputation approach is more appealing as it can be easily implemented for random intercept and slope models and is computationally faster than the fixed cluster imputation approach. Therefore, in this paper we will only study the theoretical and empirical properties of the multilevel imputation approach, although we return to fixed cluster imputation in the discussion section.

4.1 JM-MLMM

Similarly to the approach for longitudinal data, JM-MLMM uses a joint LMM for all incomplete variables. However, the formulation of the LMM will differ with respect to whether the incomplete variables are associated with random slopes or fixed effects. For example, when covariate $\boldsymbol{x}_1$ and outcome $\boldsymbol{y}$ are incomplete the following JM-MLMM model is assumed for the incomplete variables:

$$\begin{pmatrix} \boldsymbol{x}_{1i} \\ \boldsymbol{y}_i \end{pmatrix} |\boldsymbol{x}_{2i}\Bigg) = \begin{pmatrix} \alpha_{0(x_1)} + \alpha_{1(x_1)}\boldsymbol{x}_{2i} + \boldsymbol{a}_{0(x_1)i} + \boldsymbol{a}_{2(x_1)i}\boldsymbol{x}_{2i} + \boldsymbol{\varepsilon}_{1i} \\ \alpha_{0(y)} + \alpha_{1(y)}\boldsymbol{x}_{2i} + \boldsymbol{a}_{0(y)i} + \boldsymbol{a}_{1(y)i}\boldsymbol{x}_{2i} + \boldsymbol{\varepsilon}_{2i} \end{pmatrix}$$

Thus, similarly to the longitudinal settings [Appendix A.2], there is compatibility with the analysis model. As the above imputation model is similar to the longitudinal case (with $\boldsymbol{t}$ replaced by $\boldsymbol{x}_2$), we do not consider it further. However, when covariate $\boldsymbol{x}_2$ and outcome $\boldsymbol{y}$ are incomplete (scenario (iv)) the joint model for the incomplete variables using JM-MLMM is

$$\begin{pmatrix} \boldsymbol{x}_{2i} \\ \boldsymbol{y}_i \end{pmatrix} |\boldsymbol{x}_{1i}\Bigg) = \begin{pmatrix} \alpha_{0(x_2)} + \alpha_{1(x_2)}\boldsymbol{x}_{1i} + \boldsymbol{a}_{0(x_2)i} + \boldsymbol{\varepsilon}_{3i} \\ \alpha_{0(y)} + \alpha_{1(y)}\boldsymbol{x}_{1i} + \boldsymbol{a}_{0(y)i} + \boldsymbol{\varepsilon}_{4i} \end{pmatrix}$$

This JM-MLMM imputation model does not accommodate the random slope for the incomplete variable and is therefore incompatible with the analysis model [the proof is omitted as it is similar to the proof provided for the incompatibility of the JM-FJ model]. Similarly to the longitudinal setting, it can be shown that when either $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$ contains missing values but the outcome is fully observed (scenario (iii)) JM-MLMM would be compatible with the analysis model (2).

4.2 JM-FJ

The JM-FJ model for cluster-correlated data assumes the following joint model for $(\boldsymbol{y}, \boldsymbol{x}_1, \boldsymbol{x}_2)$ irrespective of missing data in $\boldsymbol{y}$, $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$

$$\begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{x}_{1i} \\ \boldsymbol{x}_{2i} \end{pmatrix} = \begin{pmatrix} \beta_{0(y)} + \boldsymbol{b}_{(y)0i} + \boldsymbol{\varepsilon}_{(y)i} \\ \beta_{0(x_1)} + \boldsymbol{b}_{(x_1)0i} + \boldsymbol{\varepsilon}_{(x_1)i} \\ \beta_{0(x_2)} + \boldsymbol{b}_{(x_2)0i} + \boldsymbol{\varepsilon}_{(x_2)i} \end{pmatrix} \tag{10}$$

This article is protected by copyright. All rights reserved.

Author Manuscript

Similarly to JM-FJ for longitudinal data (see section 3.3) it can be shown that the joint distribution implied by JM-FJ (10) is not compatible with the analysis model (2) in either scenario (iii) or (iv).

4.3 JM-SMC

As for the analysis model (1), by construction the JM-SMC approach will also be compatible with analysis model (2) irrespective of whether missing data is in the outcome or covariates.

4.4 FCS-LMM

This method uses identical imputation models to those under JM-MLMM (see section 4.1) with only one variable considered missing at a given iteration, and is compatible with the analysis model (2) irrespective of whether covariate(s) and/or the outcome are incomplete (i.e., for scenario (iii) and (iv)) if the compatibility condition (derived in 3.2) is satisfied.

4.5 FCS-LMM-het

The imputation model followed by FCS-LMM-het in the clustered data setting will be compatible with the substantive model of interest under both scenarios (iii) and (iv). The proof is similar to that provided in section 3.7 and is given in Appendix A.4.

In summary, considering all of the above methods for cluster-correlated data, we anticipated that the JM-FJ and both JM-MLMM and JM-FJ would provide biased estimates of the variance components under scenario (iii) and (iv), respectively.

5 Simulation study

In this section we describe the simulation studies that were used to assess the relative performance of the MI methods described in Sections 3 and 4 in the settings of longitudinal and clustered data. Our simulation studies are based on data from the kindergarten (K) cohort of children in LSAC (n=4983), who were aged 4-5 years when recruited in 2004. LSAC is a nationally representative study that examines the development and wellbeing of Australian children. Following recruitment, data have been collected every two years (referred to as waves of data collection) using face-to-face interviews, questionnaires and direct anthropometric measurements. The study is ongoing with six waves of data currently available. The detailed study procedure has been described elsewhere (LSAC). Here we consider two target analyses: (a) a longitudinal example: association between BMI-z score and QoL in children over time and (b) a cluster example: whether BMI-z score at wave 5 predicts the QoL at wave 6 after accounting for clustering by neighbourhood. Specifically, for analysis (a) we fitted a model similar to model (1) with QoL as a time-varying outcome, BMI-z score a time-varying covariate and age (in years) of the child as the time variable, with child-specific random intercepts and time-slopes. For analysis (b) we fitted a model similar to (2) with child QoL at wave 6 as the outcome ($\boldsymbol{y}$), child BMI-z score at wave 5 as the exposure of interest ($\boldsymbol{x}_2$) and socio-economic index for areas (SEIFA) as a covariate, with both fixed and random effects for area ($\boldsymbol{x}_1$). The missingness patterns among these variables in the LSAC dataset have been described elsewhere (Huque et al., 2018).

5.1 Longitudinal data

For the longitudinal example, we generated 1000 datasets of 5000 children assessed at 6 waves of follow-up. Three covariates at baseline: mother's education, language spoken at home and family socio-economic position; as well as three time-dependent variables: age, BMI z-score and the outcome, QoL for each child were generated. The details of the simulation setup are given below:

This article is protected by copyright. All rights reserved.

Author Manuscript

1. Whether English is the main language spoken at home (hlang) and maternal education (medu: whether or not completed year 12) for each child were generated using binomial distributions with probabilities 0.9 and 0.6 respectively.

2. The household socio-economic position (hsep) at baseline was generated using the following regression model:

$$\text{hsep}_i = -0.8 + 1.0 \times \text{medu}_i + 0.2 \times \text{hlang}_i + \nu_i, \qquad i = 1, 2, ..., 5000.$$

   where $\nu_i \sim N(0, 0.9^2)$.

3. Child age in years (cage) for the i^{th} child in the j^{th} wave (cage_{ij}) was generated according to the following model

$$\text{cage}_{ij} = \frac{1}{12}\{48 + (\text{wave}_{ij} - 1) \times 24 + \vartheta_i\} + \upsilon_{ij}, \qquad j = 1, 2, ..., 6.$$

   where $\vartheta_i = N(11, 1.5^2)$, is the distribution of age (in months) of the participant at the recruitment and $\upsilon_{ij} = N(0, 2^2)$ is the random variation in age at the time of assessment.

4. The time-varying exposure, cbmi_{ij} was then generated using the LMM

$$\text{cbmi}_{ij} = \gamma_0 + \gamma_1 \text{cage}_{ij} + \boldsymbol{u}_{0i} + \boldsymbol{u}_{1i}\text{cage}_{ij} + \Upsilon_{ij},$$

   where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ is the vector of fixed-effects, $\boldsymbol{u}_i = (\boldsymbol{u}_{0i}, \boldsymbol{u}_{1i}) \sim N(\mathbf{0}, \boldsymbol{D})$ denotes the random effects vector with the following specification for the parameters: $\boldsymbol{\gamma} = (-0.60, 0.10)^{\mathrm{T}}$, $\boldsymbol{D} = \begin{pmatrix} D_{00} & D_{01} \\ D_{10} & D_{11} \end{pmatrix} = \begin{pmatrix} 0.49 & -0.015 \\ -0.015 & 0.005 \end{pmatrix}$, where $D_{00} = \text{var}(\boldsymbol{u}_{0i})$, $D_{01} = \text{cov}(\boldsymbol{u}_{0i}, \boldsymbol{u}_{1i})$, $D_{11} = \text{var}(\boldsymbol{u}_{1i})$, and $\Upsilon_i = (\Upsilon_{i1}, \Upsilon_{i2}, ..., \Upsilon_{in_i}) \sim N(0, 0.5^2)$.

5. Finally, the continuous outcome variable, child QoL, cqol_{ij} was generated according to

$$\text{cqol}_{ij} = \beta_0 + \beta_1 \text{cbmi}_{ij} + \beta_2 \text{cage}_{ij} + \boldsymbol{b}_{0i} + \boldsymbol{b}_{1i}\text{cage}_{ij} + \boldsymbol{\varepsilon}_{ij},$$

   where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ is the vector of fixed-effects, $\boldsymbol{b}_i = (\boldsymbol{b}_{0i}, \boldsymbol{b}_{1i}) \sim N(\mathbf{0}, \boldsymbol{G})$ denotes the random effects vector. We set $\boldsymbol{\beta} = (1.00, -0.20, -0.10)^{\mathrm{T}}$, $\boldsymbol{G} = \begin{pmatrix} 0.36 & -0.012 \\ -0.012 & 0.004 \end{pmatrix}$ and residual error variance, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{in_i}) \sim N(0, 0.66^2)$.

All of the above parameter values were based on the LSAC data.

For each simulated dataset we considered two scenarios where (i) only the exposure of interest (cbmi) and (ii) both the exposure of interest (cbmi) and the outcome (cqol) were subject to missingness at each wave under an MAR mechanism. Specifically we used the following models to create missing data in cbmi and cqol, respectively

$$\begin{aligned} \text{logit}\{\Pr(R_{1ij} = 1)\} &= \theta_1 + \theta_2 \text{cqol}_{ij} + \theta_3 \text{cage}_{ij} \\ \text{logit}\{\Pr(R_{2ij} = 1)\} &= \theta_4 + \theta_5 \text{cage}_{ij} + \theta_6 \text{hsep}_{ij} \end{aligned}$$

where $R_{1ij} = 1(R_{2ij} = 1)$ if $\text{cbmi}_{ij}(\text{cqol}_{ij})$ is observed and 0 if missing. The coefficients $\boldsymbol{\theta} = (\theta_1, ..., \theta_6)^{\mathrm{T}}$ were chosen to ensure approximately 30% of the exposure (cbmi) and outcome (cqol) were missing.

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This article is protected by copyright. All rights reserved.

Author Manuscript

Author Manuscript

5.2 Clustered data

In order to evaluate the performance of the above MI methods in clustered settings, we generated 1000 datasets, each with eight variables: area identification number, socio-economic status for areas (SEIFA), mother's education (medu), language spoken at home (hlang), family socio-economic position (hsep), child sex (csex), BMI z-score (cbmi) and QoL (cqol). We considered 300 areas (clusters), where the number of children in each area varied between 2 to 25. Our simulated dataset mimicked the LSAC dataset not only in terms of cluster size and the number of clusters, but also with regards to the relationship between the covariates. The analysis of interest was whether the relationship between child BMI z-score at wave 5 and QoL at wave 6 varied across all areas. In all of the simulated datasets variables were simulated in a sequential manner as follows:

1. Sex (csex), English language background (hlang) and mother's education (medu: whether or not completed year 12) for each child were generated using binomial distributions with probabilities 0.5, 0.9 and 0.6 respectively.

2. Child age in years at wave 5 was generated using the following model

   $$\text{cage}_{ij} = \frac{1}{12}\{168 + \vartheta_{ij}\} \qquad j = 2, 3, ..., 25.i = 1, 2, ...300$$

   where $\vartheta_{ij} = N(11, 1.5^2)$ is the distribution of age (in months) of the j^{th} child at recruitment from area i.

3. The main exposure variable of interest, cbmi was generated based on child's age and sex using the following linear regression model

   $$\text{cbmi}_{ij} = (-1.0 + \boldsymbol{d}_{0i}) + 0.11 * \text{cage}_{ij} + 0.05 * \text{csex}_{ij} + \psi_{ij},$$

   where $\psi_{ij} \sim N(0, 1)$ and $\boldsymbol{d}_{0i} \sim N(0, 0.15^2)$

4. SEIFA at each area was generated as a standard normal variable.

5. Family socio-economic position (hsep) was generated based on SEIFA, mother's education and language using the following linear regression model

   $$\text{hsep}_{ij} = -4.7 + 0.8 * \text{medu}_{ij} + 0.01 * \text{SEIFA}_i + 0.01 * \text{hlang}_{ij} + \phi_{ij}$$

   where $\phi_{ij} \sim N(0, 0.9^2)$.

6. Finally the outcome, cqol score, was generated using the LMM

   $$\text{cqol}_{ij} = (0.05 + \boldsymbol{b}_{0i}) + (-0.2 + \boldsymbol{b}_{1i}) * \text{cbmi}_{ij} + 0.25 * \text{SEIFA}_i + e_{ij}$$

   with $e_{ij} \sim N(0, 0.9^2)$, $\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{G})$ and $\boldsymbol{G} = \begin{pmatrix} 0.16 & 0.00 \\ 0.00 & 0.04 \end{pmatrix}$.

   All of the above parameter values were based on LSAC data. The exception was that we slightly inflated the magnitude of the regression and variance-component parameters in the outcome model in order to accentuate the differences in the estimated parameters from the MI methods.

   For each simulated dataset we considered two scenarios: (iii) only the exposure of interest cbmi and (iv) both the exposure cbmi and outcome cqol were missing under an MAR mechanism. Specifically, we fitted the following models to create missing data in cbmi and cqol, respectively

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This article is protected by copyright. All rights reserved.

Author Manuscript

$$\begin{aligned}\text{logit}\{\Pr(R_{3ij}=1)\} &= -2.2+1.0*\text{cqol}_{ij}+0.2*\text{SEIFA}_i-0.2*\text{hsep}_{ij}\\ \text{logit}\{\Pr(R_{4ij}=1)\} &= -2.5+0.2*\text{SEIFA}_i-0.3*\text{hsep}_{ij}\end{aligned}$$

where $R_{3ij}=1(R_{4ij}=1)$ if cbmi(cqol) is observed and 0 if missing.

6 Performance of the MI method

We applied all the imputation methods described in section 3 and 4 to the simulated and LSAC datasets. In light of the seven main choices for the specification of multiple imputation method namely i) the MAR assumption, ii) form of the imputation model, (iii) set of variables included in the imputation model, iv) passive imputation v) order of the variables vi) number of iterations and vii) number of multiply imputed datasets (Van Buuren and Groothuis-Oudshoorn, 2011), we generated data under MAR and included the same set of predictors across all the imputation methods. Specifically, we included socio-economic position as an auxiliary variable, in addition to all analysis variables, and considered the same order of imputation variables for all the methods (if applicable). Thirty imputations were generated for each approach to limit Monte Carlo (imputation-related) error for the regression coefficient of interest to approximately 5 percent of its standard error. However, for each method, we set the number of burn-in iterations and number of between imputations to the default values of current software implementations and finally the form of the imputation models varies according to the specific imputation method. We compared estimated regression coefficients, standard errors (both average of the model-based and the empirical standard error) and variance-component estimates from the various imputation approaches and an available data analysis, which excluded records with missing data in any analysis variable. Bias and coverage probability of the estimated regression coefficients from each of the approaches and from an available data analysis compared to the values used to generate the data are also presented. In each case, the sampling properties of the estimators were estimated from the 1000 simulated datasets.

6.1 Simulation results

The simulation results for the longitudinal example with missing values under scenarios (i) and (ii) across the 1000 simulated datasets are displayed in Table 1 and 2, respectively. It is clear from Table 1 and 2 that the available data analysis resulted in biased estimation of the regression coefficients and variance components along with inadequate coverage probabilities.

All of the MI approaches except JM-FJ provided similar estimation of regression parameters and coverage probabilities in both scenarios. Slight under-coverage of the regression parameters was obtained from JM-FJ and JM-SMC, which assume homoscedastic variances. However, somewhat contrasting results were obtained when imputed assuming heteroscedastic covariance matrices for both of these methods. The JM-FJ was more biased and led to underestimation of coverage probabilities while JM-SMC performed better compared with its homoscedastic counterpart.

All of the MI methods except JM-FJ-het provided unbiased estimates of the variance components in the longitudinal setting when only the covariate contained missing values (scenario (i)). However, greater differences were observed across the imputation methods for the estimation of the variance components when both covariate and outcome contained missing values (scenario (ii)). In this scenario (ii) (Table 2), large biases in the variance associated with random slopes were obtained for the JM-MLMM, JM-FJ, JM-FJ-het and FCS-LMM approaches. ~~The large bias in the variance associated with random slopes with the JM-MLMM and FCS-LMM approaches is likely an artefact of dividing by a population value that is close to zero, hence we are hesitant to emphasize this finding~~.

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This article is protected by copyright. All rights reserved.

Author Manuscript

Following a reviewer's suggestions, we also evaluated the performance of these methods in the case of smaller samples with 1000 individuals followed for 5 consecutive period under both scenarios (i) and (ii). The results, displayed in Tables B1 and B2 in the Appendix B, are qualitatively similar to those with the larger sample size under scenario (i). But under scenario (ii) large biases associated with random slopes variance estimates were obtained for all the methods except JM-MVN, FCS-standard and JM-SMC approaches. Among the MI methods, JM-MVN and FCS-Standard provided the least biased estimates for the fixed effects and variance components. The estimated coverage probabilities for both of these methods were very close to the nominal value of 0.95, in both scenarios. Among the LMM-based imputation approaches, ~~FCS-LMM-het and~~ JM-SMC-het provided the best performance for estimating regression parameters and variance components.

The simulation results for clustered data with missing values under scenarios (iii) and (iv) across the 1000 simulated datasets are displayed in Tables 3 and 4, respectively. Similarly to the longitudinal setting, in the clustered data setting, the available data analysis resulted in biased estimation of the regression coefficients and variance components, and inadequate coverage probabilities. All of the MI approaches provided similar estimates of the regression coefficients and their estimated coverage probabilities were very close to the nominal value of 0.95 for both scenarios. Slight under-coverage of the confidence interval for the regression coefficient of cbmi was obtained from JM-FJ, JM-FJ-het and JM-MLMM especially under scenario (iv). Somewhat greater differences were observed across the imputation methods for the estimation of variance components both in scenario (iii) and (iv). In scenario (iii), i.e. when only the covariate with the random effect contained missing values, JM-FJ and JM-FJ-het resulted in biased estimation of the random slope variances. On the other hand, in scenario (iv) JM-FJ, JM-FJ-het and JM-MLMM all produced biased estimation of the random slope variances. In both scenarios, JM-SMC, FCS-LMM and FCS-LMM-het produced unbiased estimates of the regression and variance component parameters.

As with the longitudinal data we also evaluated the performance of these methods under scenario (iii) and (iv) using a relatively small number of clusters (n=100) with smaller cluster sizes (each cluster contained between 2 and 10 observations randomly). The results are displayed in Table B3 and B4, respectively. All of the MI methods except JM-FJ in both scenarios and JM-MLMM in scenario (iV) provided slight under-coverage of the confidence interval. Large biases in the estimation of the random slope parameters were observed for both FCS-LMM and FCS-LMM-het, especially in scenario (iv), leaving JM-SMC as the best methods when the number of cluster in the sample is small.

6.2 Application to the LSAC data

The results for the analysis models (a) and (b) applied to the LSAC data are given in Tables 5 and 6 respectively. Available data analysis provides slightly lower estimates of the regression coefficients in the case for analysis model (b) compared with the estimates from MI methods. However, for analysis model (a) the estimated regression coefficients from the available data analysis are very similar to those from the MI approaches. These results are in line with those seen in the simulation study. However, JM-FJ in analysis model (a) and both JM-FJ and JM-MLMM in analysis model (b) produced lower estimates of the variance components than the other MI approaches.

7 Discussion

LMMs are frequently used in the analysis of longitudinal and clustered data in order to account for within-individual and within-cluster correlation, respectively. Although several MI methods are available for imputing missing values in longitudinal and cluster-correlated data in the current software, little guidance is available on which is the most appropriate method. The comparison of MI methods for the analysis of correlated data using LMM with random intercepts and slopes in the context of compatibility, a prerequisite of valid MI, is very limited in the literature. In the current paper, we compared seven different MI methods (JM-MVN, JM-MLMM, JM-FJ, JM-SMC, FCS-Standard, FCS-LMM, FCS-LMM-het) for

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This article is protected by copyright. All rights reserved.

handling missing values in longitudinal and clustered data in the context of fitting LMM with both random intercepts and slopes. We derived expressions for each of the MI methods to examine the compatibility of these MI methods with a LMM that include both random intercepts and slopes. We showed that compatible imputation and analysis models resulted in consistent estimation of both regression parameters and variance components via simulation. We have summarized our results in Table 7.

The results from our theoretical exploration revealed that the relative performance of the MI methods may be expected to vary according to whether the incomplete covariate has fixed or random effects and to the missingnesss in the outcome variable. Specifically, we showed that JM-MVN and FCS-Standard approaches are compatible with the LMM in the context of longitudinal data if measurements occur at the same time-points for all individuals. We also showed that JM-MLMM is compatible with, but that JM-FJ is incompatible with the analysis model of a LMM with random intercepts and slopes. Both the FCS-LMM and FCS-LMM-het methods are compatible with a LMM with random intercepts and slopes. Our comparison also revealed that the newly available substantive model compatible joint modeling (JM-SMC) approach holds great promise for the imputation of longitudinal data. Our simulation study supported our theoretical results. We observed, however, that JM-FJ-het provided sub-optimal performance, especially in the case of longitudinal data, which might be due to a small number of individuals per cluster (observation per individual) in our example, as shown in Audigier et al. (2018). We also observed that JM-SMC-het provided better estimates for the regression parameters and coverage than JM-SMC, apparently because subject-specific associations were better estimated under the heteroscedastic covariance matrices.

Our results regarding clustered data were similar to those for longitudinal data with a couple of exceptions. We found JM-MLMM was compatible with a LMM with a random intercepts and slopes analysis model if only the covariate contains missing data. The JM-MLMM, however, became non-compatible with a LMM with random intercepts and slopes if both the outcome and random-slope covariate contained missing data. Along with others (Enders et al., 2016), we noted that fixed effect imputation methods are computationally expensive particularly with a large number of clusters, hence may not be very useful in practice. In general, our findings are consistent with those of (Enders et al., 2016) who showed that JM-FJ and JM-MLMM produce biased estimation of the variance components while the FCS-LMM-het approach provided consistent estimates in the context of clustered data. Some of our theoretical results extend the results obtained by Resche-Rigon and White (Resche-Rigon and White, 2016) who considered a LMM with only a random intercept.

It is always difficult to draw general conclusions from a single simulation study, but we believe this study provided a good setting for a comparison of MI methods with both theoretical and empirical evaluation. The simulations were designed to represent real world data with a moderate amount of missingness under MAR. Undoubtedly, future simulation studies and further exploration of methods will be useful in a number of ways. In this study, we restricted our attention to data that are MAR. Often longitudinal data does not satisfy the MAR assumption. In general, the MAR assumption cannot be tested but various sensitivity analysis methods (e.g., selection models, pattern-mixture model and NARFCS) are proposed in very specialized context and no such analysis methods is currently available for the context when both longitudinal covariates and outcomes are missing. Forexample, pattern-mixture models are available in the context of longitudinal outcomes but not for the context when both longitudinal covariates and outcomes are missing. The NARFCS approach, arising from the pattern-mixture paradigm, has been developed recently to handle multivariable missingness in cross-sectional settings(Tompsett et al., 2018; Moreno-Betancur et al., 2017; Leacy, 2016) and these could in principle be applied for longitudinal data in the wide format or with the cluster-indicator method in the scenarios we explored. However, these methods have not yet been extended to the context of multilevel imputation models for general multivariable missingness in longitudinal unbalanced data or clustered data (linear mixed models). Hence there were no obvious methods to add to our evaluation. In order to simplify the theoretical calculations and avoid mis-specification of the imputation models we restricted our comparisons to models and methods that assume normality. Although there has been some discussion of compatibility for non-normal data in the context of general location models, such models are only available for single level data (Seaman and Hughes, 2018). The study of compatibility

This article is protected by copyright. All rights reserved.

Author Manuscript

of multilevel models that include non-normal data is beyond the scope of the present paper as Gaussian random effects are usually assumed in the proposed models and in the available software implementation. However, our results for MI involving normal variables might also hold for non-normal data. We had previously shown that both JM-MVN and FCS-standard showed good performance in the context of imputing binary variables (Huque et al., 2018). Quartagno et al. recently showed that the JM-SMC and FCS-standard methods performed equally well in the context of imputing non-normal data (Quartagno and Carpenter, 2019).

In summary, we found that if measurements occur at the same time-points for all individuals in longitudinal studies, the JM-MVN and FCS-Standard approaches may be the best approaches for imputing longitudinal data. We also found that LMM-based approaches (JM-MLMM, JM-SMC-het, FCS-LMM-het, FCS-LMM) can be used if measurement doesn't occur at the same time points or the imputation model struggles to converge due to many repeated measurements. In the clustered data setting, we recommend using the LMM-based approaches JM-SMC, FCS-LMM or FCS-LMM-het to handling missing data as they performed well in the estimation of regression parameters and variance components. Although multilevel imputation models are slightly more complex compared with standard cross-sectional imputation methods and require specialized software, our comparison revealed that they are a reasonable choice for imputing missing covariate and outcome data where the analysis of interest is a linear mixed effect model with random intercepts and slopes.

Acknowledgements

This work was supported by funding from the National Health and Medical Research Council: Project grant ID# 1102468, Career Development Fellowship ID#1127984 (KJL), Senior Research Fellowship ID# 1104975 (JAS), and Centre of Research Excellence grant ID#1035261, for the Victorian Centre for Biostatistics (ViCBiostat). Research at the Murdoch Childrens Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program.

References

Tihomir Asparouhov and Bengt Muthén. Multiple imputation with mplus. *MPlus Web Notes*, 2010.

Vincent Audigier, Ian R White, Shahab Jolani, Thomas Debray, Matteo Quartagno, James Carpenter, Stef van Buuren, and Matthieu Resche-Rigon. Multiple imputation for multilevel data with continuous and binary variables. *Statistical Science*, 33(2):160–183, 2018.

Craig K Enders, Stephen A Mistler, and Brian T Keller. Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological methods*, 21(2):222–240, 2016.

Craig K Enders, Brian T Keller, and Roy Levy. A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological methods*, 2017.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439840955. URL `https://books.google.com.au/books?id=ZXL6AQAAQBAJ`.

Harvey Goldstein, James Carpenter, Michael G Kenward, and Kate A Levin. Multilevel models with multivariate mixed response types. *Statistical Modelling*, 9(3):173–197, 2009.

Harvey Goldstein, James R Carpenter, and William J Browne. Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(2):553–564, 2014.

Simon Grund, Oliver Lüdtke, and Alexander Robitzsch. Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. *Behavior Research Methods*, 48(2): 640–649, 2016.

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

Rachael A Hughes, Ian R White, Shaun R Seaman, James R Carpenter, Kate Tilling, and Jonathan AC Sterne. Joint modelling rationale for chained equations. *BMC medical research methodology*, 14(1): 28, 2014.

Md Hamidul Huque, John B Carlin, Julie A Simpson, and Katherine J Lee. A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology*, xx(xx):xx–xx, 2018.

Nan M Laird. Missing data in longitudinal studies. *Statistics in medicine*, 7(1-2):305–315, 1988.

FP Leacy. *Multiple imputation under missing not at random assumptions via fully conditional specification [dissertation]*. PhD thesis, 2016.

Jingchen Liu, Andrew Gelman, Jennifer Hill, Yu-Sung Su, and Jonathan Kropko. On the stationary distribution of iterative imputations. *Biometrika*, 101(1):155–173, 2014.

LSAC. Technical report.

Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, pages 538–558, 1994.

M Moreno-Betancur, FP Leacy, D Tompsett, and I White. mice: The narfcs procedure for sensitivity analyses, 2017. URL `https://github.com/moreno-betancur/NARFCS`.

M Quartagno and JR Carpenter. Multiple imputation for ipd meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in medicine*, 35(17):2938–2954, 2016.

Matteo Quartagno and James R Carpenter. Multiple imputation for discrete data: Evaluation of the joint latent normal model. *Biometrical Journal*, 2019.

Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, and Peter Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96, 2001.

Jerome P Reiter, Trivellore E Raghunathan, and Satkartar K Kinney. The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, 32(2):143, 2006.

Matthieu Resche-Rigon and Ian R White. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical methods in medical research*, page 0962280216666564, 2016.

Panteha Hayati Rezvan, Katherine J Lee, and Julie A Simpson. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC medical research methodology*, 15(1):30, 2015.

D. B. Rubin. *Multiple imputation for nonresponse in surveys/Donald B. Rubin*. Wiley, c1987, New York, 1987.

Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

Joseph L Schafer and Recai M Yucel. Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of computational and Graphical Statistics*, 11(2):437–457, 2002.

Shaun R Seaman and Rachael A Hughes. Relative efficiency of joint-model and full-conditional-specification multiple imputation when conditional models are compatible: The general location model. *Statistical methods in medical research*, 27(6):1603–1614, 2018.

J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, and M. G. Kenward. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, 338, 2009.

Daniel Mark Tompsett, Finbarr Leacy, Margarita Moreno-Betancur, Jon Heron, and Ian R White. On the use of the not-at-random fully conditional specification (narfcs) procedure in practice. *Statistics in medicine*, 37(15):2338–2353, 2018.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3):1–67, 2011.

This article is protected by copyright. All rights reserved.

Stef Van Buuren, Jaap PL Brand, CGM Groothuis-Oudshoorn, and Donald B Rubin. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76(12): 1049–1064, 2006.

Stef Van Buuren et al. Multiple imputation of multilevel data. *Handbook of advanced multilevel analysis*, pages 173–196, 2011.

Recai M Yucel. Random covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical modelling*, 11(4):351–370, 2011.

Enxu Zhao and Recai M Yucel. Performance of sequential imputation method in multilevel applications. In *American Statistical Association Proceedings of the Survey Research Methods Section*, pages 2800–2810, 2009.

Jian Zhu and Trivellore E Raghunathan. Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association*, 110(511):1112–1124, 2015.

Author Manuscript

This article is protected by copyright. All rights reserved.

Table 1 Simulation results for the analysis of longitudinal data using simulation scenario (i), i.e., scenario with data missing in the covariate only

Regression Parameters	True value	Available data	JM-MVN	JM-MLMM	JM-FJ	JM-FJ-het	JM-SMC	JM-SMC-het	FCS-standard	FCS-LMM	FCS-LMM-het
cbmi, $\widehat{\beta_1}$	-0.200	-0.184	-0.200	-0.204	-0.194	-0.186	-0.195	-0.197	-0.199	-0.204	-0.204
$\widehat{\beta_1}$ rbias (%)		0.078	0.002	0.020	0.029	0.068	0.027	0.014	0.003	0.021	0.019
Model SE		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
Empirical SE		0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008
95% Coverage		0.463	0.945	0.921	0.895	0.644	0.891	0.941	0.950	0.919	0.926
cage, $\widehat{\beta_2}$	-0.100	-0.090	-0.100	-0.100	-0.101	-0.102	-0.101	-0.101	-0.100	-0.100	-0.100
$\widehat{\beta_1}$ rbias (%)		0.105	0.001	0.003	0.009	0.016	0.010	0.007	<0.001	0.003	0.003
Model SE		0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Empirical SE		0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
95% Coverage		0.000	0.948	0.947	0.926	0.853	0.923	0.937	0.951	0.948	0.949
Variance components											
G_{00} est	0.360	0.335	0.361	0.364	0.360	0.361	0.361	0.360	0.360	0.364	0.363
G_{00}rbias (%)		0.071	0.002	0.010	<0.001	0.002	0.003	0.001	<0.001	0.010	0.009
G_{01} est	-0.012	-0.011	-0.012	-0.012	-0.012	-0.012	-0.012	-0.012	-0.012	-0.012	-0.012
G_{01} rbias (%)		0.044	<0.001	0.028	0.012	0.002	0.022	0.022	0.005	0.028	0.024
G_{11} est	0.004	0.003	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
G_{11} rbias(%)		0.170	0.003	0.013	0.008	0.003	0.013	0.010	<0.001	0.013	0.011
Residual error, $\widehat{\sigma}_{\varepsilon}^2$	0.436	0.414	0.436	0.434	0.436	0.437	0.436	0.435	0.436	0.434	0.434
$\widehat{\sigma}_{\varepsilon}$ rbias (%)		0.049	<0.001	0.003	0.001	0.003	<0.001	0.002	<0.001	0.003	0.003

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

Table 2 Simulation results for the analysis of longitudinal data using simulation scenario (ii), i.e., scenario with data missing in both covariate and outcome

Regression Parameters	True value	Available data	JM-MVN	JM-MLMM	JM-FJ	JM-FJ-het	JM-SMC	JM-SMC-het	FCS-standard	FCS-LMM	FCS-LMM-het
cbmi, $\widehat{\beta_1}$	-0.200	-0.183	-0.198	-0.198	-0.197	-0.185	-0.195	-0.197	-0.198	-0.202	-0.202
$\widehat{\beta_1}$ rbias (%)		0.083	0.011	0.010	0.015	0.075	0.025	-0.013	0.011	0.011	0.011
Model SE		0.008	0.010	0.010	0.010	0.010	0.009	0.009	0.010	0.009	0.009
Empirical SE		0.008	0.010	0.010	0.010	0.009	0.009	0.009	0.010	0.009	0.009
95% Coverage		0.510	0.940	0.936	0.923	0.697	0.913	0.936	0.943	0.939	0.934
cage, $\widehat{\beta_2}$	-0.100	-0.088	-0.100	-0.100	-0.101	-0.102	-0.101	-0.101	-0.100	-0.100	-0.100
$\widehat{\beta_1}$ rbias (%)		0.119	0.003	0.002	0.005	0.017	0.009	0.007	0.003	0.002	0.002
Model SE		0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Empirical SE		0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
95% Coverage		0.000	0.948	0.950	0.933	0.867	0.915	0.931	0.943	0.959	0.966
Variance components											
G_{00} est	0.360	0.333	0.359	0.375	0.345	0.378	0.360	0.359	0.360	0.378	0.377
G_{00}rbias (%)		0.075	0.003	0.041	0.043	0.051	0.001	0.003	0.001	0.051	0.046
G_{01} est	-0.012	-0.011	-0.012	-0.014	-0.002	-0.007	-0.012	-0.012	-0.012	-0.015	-0.013
G_{01} rbias (%)		0.050	0.012	0.207	0.818	0.405	0.012	0.006	0.001	0.227	0.099
G_{11} est	0.004	0.003	0.004	0.004	0.002	0.003	0.004	0.004	0.004	0.004	0.004
G_{11} rbias(%)		0.176	0.003	0.101	0.494	0.405	0.011	0.008	0.036	0.107	0.031
Residual error, $\widehat{\sigma}_{\varepsilon}^2$	0.436	0.415	0.438	0.436	0.463	0.487	0.436	0.435	0.438	0.432	0.435
$\widehat{\sigma}_{\varepsilon}$ rbias (%)		0.047	0.005	0.006	0.063	0.117	0.001	0.001	0.005	0.008	0.001

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This article is protected by copyright. All rights reserved.

Table 3 Simulation results for the analysis of clustered data using simulation scenario (iii),i.e., scenario with data missing in the covariate only

Regression Parameters	True value	Available data	JM-MLMM	JM-FJ	JM-FJ-het	JM-SMC	JM-SMC-het	FCS LMM	FCS-LMM-het
SEIFA, $\widehat{\gamma}_1$	0.250	0.192	0.247	0.250	0.192	0.250	0.250	0.247	0.249
$\widehat{\gamma}_1$ rbias (%)		0.232	0.011	<0.001	0.232	0.001	0.001	0.011	0.004
Model SE		0.028	0.029	0.029	0.028	0.029	0.029	0.029	0.029
Empirical SE		0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028
95% Coverage		0.444	0.954	0.950	0.450	0.947	0.949	0.954	0.950
cbmi, $\widehat{\gamma}_2$	-0.200	-0.177	-0.196	-0.205	-0.177	-0.200	-0.200	-0.196	-0.198
$\widehat{\gamma}_2$ rbias (%)		0.115	0.019	0.027	0.001	0.001	0.001	0.020	0.009
Model SE		0.020	0.023	0.021	0.021	0.023	0.023	0.023	0.023
Empirical SE		0.021	0.023	0.024	0.021	0.024	0.024	0.023	0.023
95% Coverage		0.793	0.958	0.913	0.796	0.946	0.947	0.953	0.945
Variance components									
G_{00} est	0.40	0.346	0.396	0.404	0.348	0.399	0.399	0.396	0.397
G_{00}rbias (%)		0.136	0.011	0.010	0.115	0.002	0.002	0.011	0.007
G_{11} est	0.20	0.175	0.197	0.135	0.176	0.192	0.192	0.197	0.185
G_{11} rbias(%)		0.123	0.014	0.325	0.232	0.039	0.038	0.016	0.074
Residual error, $\widehat{\sigma}_e$	0.90	0.847	0.900	0.911	0.847	0.901	0.901	0.900	0.903
$\widehat{\sigma}_e$ rbias (%)		0.058	<0.001	0.012	0.058	0.001	0.001	0.001	0.003

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

Table 4 Simulation results for the analysis of clustered data using simulation scenario (iv), i.e., senario with data missing in covariate associated with random slope and outcome

Regression Parameters	True value	Available data	JM-MLMM	JM-FJ	JM-FJ-het	JM-SMC	JM-SMC-het	FCS-LMM	FCS-LMM-het
SEIFA, $\widehat{\gamma}_1$	0.250	0.191	0.249	0.250	0.192	0.250	0.250	0.247	0.249
$\widehat{\gamma}_1$ rbias (%)		0.235	0.002	<0.001	0.232	<0.001	0.001	0.011	0.004
Model SE		0.031	0.029	0.029	0.028	0.029	0.029	0.029	0.029
Empirical SE		0.031	0.029	0.028	0.028	0.028	0.028	0.028	0.028
95% Coverage		0.508	0.952	0.950	0.450	0.948	0.948	0.953	0.950
cbmi, $\widehat{\gamma}_1$	-0.200	-0.179	-0.205	-0.205	-0.177	-0.200	-0.200	-0.196	-0.198
$\widehat{\gamma}_2$ rbias (%)		0.105	0.024	0.026	0.115	0.001	0.001	0.021	0.009
Model SE		0.024	0.021	0.021	0.021	0.023	0.023	0.023	0.023
Empirical SE		0.024	0.024	0.024	0.021	0.024	0.024	0.023	0.023
95% Coverage		0.853	0.912	0.912	0.796	0.945	0.943	0.951	0.943
Variance components									
G_{00} est	0.40	0.342	0.404	0.404	0.348	0.399	0.399	0.396	0.397
G_{00}rbias (%)		0.144	0.010	0.010	0.130	0.002	0.002	0.010	0.007
G_{11} est	0.20	0.171	0.134	0.135	0.176	0.192	0.192	0.197	0.1185
G_{11} rbias(%)		0.143	0.330	0.325	0.118	0.039	0.039	0.015	0.073
Residual error, $\widehat{\sigma}_e$	0.900	0.848	0.911	0.911	0.847	0.901	0.901	0.900	0.903
$\widehat{\sigma}_e$ rbias (%)		0.057	0.012	0.012	0.058	0.001	0.001	<0.001	0.003

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This article is protected by copyright. All rights reserved.

Table 5 LSAC data analysis for analysis model (1) i.e., longitudinal data scenario.

Regression Parameters	Available data	JM-MVN	JM-MLMM	JM-FJ	JM-FJ-het	JM-SMC	JM-SMC-het	FCS-standard	FCS-LMM	FCS-LMM-het
cbmi, $\widehat{\beta_1}$	-0.043	-0.043	-0.042	-0.047	-0.042	-0.042	-0.045	-0.044	-0.043	-0.047
se($\widehat{\beta_1}$)	0.008	0.008	0.008	0.007	0.008	0.008	0.008	0.008	0.007	0.010
95% CI	(-0.058, -0.029)	(-0.058, -0.027)	(-0.058, -0.026)	(-0.061, -0.033)	(-0.058, -0.029)	(-0.058, -0.026)	(-0.061, -0.029)	(-0.059, -0.029)	(-0.057, -0.029)	(-0.067, -0.027)
cage, $\widehat{\beta_2}$, years	-0.020	-0.022	-0.021	-0.019	-0.020	-0.021	-0.021	-0.022	-0.021	-0.018
se($\widehat{\beta_2}$)	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
95% CI	(-0.024, -0.016)	(-0.026, -0.018)	(-0.025, -0.017)	(-0.023, -0.015)	(-0.024, -0.016)	(-0.025, -0.017)	(-0.025, -0.017)	(-0.026, -0.018)	(-0.025, -0.017)	(-0.022, -0.014)
Variance components										
$\widehat{G}_{00}$	0.431	0.518	0.473	0.419	0.432	0.443	0.450	0.520	0.472	0.441
$\widehat{G}_{01}$	-0.016	-0.021	-0.020	-0.007	-0.016	-0.016	-0.017	-0.022	-0.020	-0.016
$\widehat{G}_{11}$	0.004	0.004	0.005	0.002	0.004	0.004	0.004	0.005	0.005	0.004
$\widehat{\sigma}_{\varepsilon}^2$	0.434	0.436	0.434	0.460	0.434	0.438	0.437	0.436	0.434	0.465

This article is protected by copyright. All rights reserved.

Author Manuscript

Author Manuscript

Table 6 LSAC data analysis for analysis model (2) of the cluster-correlated data

Regression Parameters	Available data	JM-MLMM	JM-FJ	JM-FJ-het	JM-SMC	JM-SMC-het	FCS-LMM	FCS-LMM-het
SEIFA,$\widehat{\gamma}_1$	0.154	0.164	0.166	0.154	0.166	0.166	0.166	0.159
se($\widehat{\gamma}_1$)	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019
95% CI	(0.117, 0.191)	(0.127, 0.201)	(0.129, 0.203)	(0.117, 0.191)	(0.129, 0.203)	(0.129, 0.203)	(0.129, 0.203)	(0.122, 0.196)
cbmi,$\widehat{\gamma}_2$	-0.071	-0.075	-0.073	-0.071	-0.078	-0.079	-0.075	-0.073
se($\widehat{\beta_2}$)	0.017	0.018	0.017	0.017	0.018	0.018	0.019	0.018
95% CI	(-0.104,-0.038)	(-0.110, -0.040)	(-0.106, -0.040)	(-0.104,-0.038)	(-0.113, -0.043)	(-0.114, -0.044)	(-0.112, -0.038)	(-0.108, -0.038)
Variance components								
$\widehat{G}_{00}$	0.029	0.035	0.034	0.029	0.031	0.031	0.035	0.030
$\widehat{G}_{11}$	0.007	0.006	0.005	0.007	0.008	0.008	0.014	0.008
$\widehat{\sigma}_e$	0.931	0.923	0.929	0.867	0.923	0.923	0.916	0.937

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This article is protected by copyright. All rights reserved.

Table 7 Summary of multiple imputation models features

Features	JM- MVN	JM- MLMM	JM- FJ	JM- SMC	FCS- standard	FCS- LMM	FCS- LMM-het
Unbalanced data	No	Yes	Yes	Yes	No	Yes	Yes
Imputation of discrete variables	continuous	continuous	latent normal	latent normal	categorical	continuous	continuous
Analysis of discrete variables	require rounding	require rounding	Yes	Yes	Yes	require rounding	require rounding
Longitudinal data							
Consistent estimation of random intercepts	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Consistent estimation of random slopes							
with incomplete covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
with incomplete covariates and outcome	Yes	No	No	Yes	Yes	No	No
Clustered data							
Consistent estimation of random intercepts		Yes	Yes	Yes		Yes	Yes
Consistent estimation of random slopes							
with incomplete covariates		No	No	Yes		No	Yes
with incomplete covariates and outcome		No	No	Yes		No	No

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This article is protected by copyright. All rights reserved.

Author Manuscript

Author Manuscript

Appendix

A Compatibility of the methods

A.1 JM-MVN

Suppose we are interested in substantive model (1). Now assume that the covariate $\boldsymbol{x}_i$ also follows the following LMM

$$\boldsymbol{x}_i|\boldsymbol{t}_i \quad = \quad \gamma_0 + \gamma_1 \boldsymbol{t}_i + \boldsymbol{u}_{0i} + \boldsymbol{u}_{1i}\boldsymbol{t}_i + \boldsymbol{\epsilon}_i, \tag{11}$$

where $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \Upsilon)$ and $\boldsymbol{u}_i = (\boldsymbol{u}_{0i}, \boldsymbol{u}_{1i}) \sim N(\mathbf{0}, \boldsymbol{D})$. As both the distribution of $(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{t}_i)$ and $(\boldsymbol{x}_i|\boldsymbol{t}_i)$ are Gaussian. The joint distribution $(\boldsymbol{y}_i, \boldsymbol{x}_i|\boldsymbol{t}_i)$ can be written as (Gelman et al., 2013)

$$\begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{x}_i \end{pmatrix} |\boldsymbol{t}_i \Bigg) = N\left(\begin{pmatrix} \beta_0 + \beta_1(\gamma_0 + \gamma_1 \boldsymbol{t}_i) + \beta_2 \boldsymbol{t}_i \\ \gamma_0 + \gamma_1 \boldsymbol{t}_i \end{pmatrix}, \begin{pmatrix} \beta_1 \Sigma_{ix} \beta_1^{\mathrm{T}} + \Sigma_{iy} & \beta_1 \Sigma_{ix} \\ \Sigma_{ix}\beta_1 & \Sigma_{ix} \end{pmatrix}\right), \tag{12}$$

where $\Sigma_{ix} = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^{\mathrm{T}} + \Upsilon$ and $\Sigma_{iy} = \boldsymbol{Z}_i \boldsymbol{G} \boldsymbol{Z}_i^{\mathrm{T}} + \Phi$ with $\boldsymbol{Z}_i = (\mathbf{1}, \boldsymbol{t}_i)^{\mathrm{T}}$. If data are collected for an equal number of visits and fixed time interval between successive visits for all individuals, then the joint distribution of $(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{t})$ can be written as

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{pmatrix} |\boldsymbol{t} \Bigg) \quad = \quad N\left(\begin{pmatrix} \beta_0 + \beta_1(\gamma_0 + \gamma_1 \boldsymbol{t}) + \beta_2 \boldsymbol{t} \\ \gamma_0 + \gamma_1 \boldsymbol{t} \end{pmatrix}, \begin{pmatrix} \beta_1 \Sigma_x \beta_1^{\mathrm{T}} + \Sigma_y & \beta_1 \Sigma_x \\ \Sigma_x \beta_1 & \Sigma_x \end{pmatrix}\right). \tag{13}$$

A.2 JM-MLMM

scenario i: If only covariates have missing data, the substantive and the imputation models for the i^{th} subject can be written as

$$(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{t}_i) \sim N(\beta_0 + \beta_1 \boldsymbol{x}_i + \beta_2 \boldsymbol{t}_i, \Sigma_{iy}) \tag{14}$$

and

$$(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{t}_i) \sim N(\beta_{0(x)} + \beta_{1(x)} \boldsymbol{y}_i + \beta_{2(x)} \boldsymbol{t}_i, \Sigma_{ix|y}) \tag{15}$$

These two models are compatible if β's, $\beta_{(x)}$'s, $\Sigma_{iy|x}$ and $\Sigma_{ix|y}$ lie on the subspace of the joint model determined by (12) (under the assumption that covariate $\boldsymbol{x}_i$ also follows a LMM (11)).

Therefore the conditional distribution $(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{t}_i)$ from the above joint distribution can be written as

$$(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{t}_i) = N\left(\Sigma_{ix|y}\left[\beta_1^{\mathrm{T}} \Sigma_{iy}^{-1}(\boldsymbol{y}_i - (\beta_0 + \beta_2 \boldsymbol{t}_i)) + \Sigma_{ix}^{-1}(\gamma_0 + \gamma_1 \boldsymbol{t}_i)\right], \Sigma_{ix|y} = (\beta_1^{\mathrm{T}} \Sigma_{iy}^{-1} \beta_1 + \Sigma_{ix}^{-1})^{-1}\right) \tag{16}$$

Now equating (15) and (16) we have:

$$\beta_{1(x)}^{\mathrm{T}} \Sigma_{ix|y}^{-1} = \beta_1^{\mathrm{T}} \Sigma_{iy}^{-1}$$

Hence the substantive model will be compatible to the imputation model under JM-MLMM if $\beta_{1(x)}^{\mathrm{T}} \Sigma_{x|y}^{-1} = \beta_1^{\mathrm{T}} \Sigma_y^{-1}$.

scenario ii: Both covariate and outcome have missing data

This article is protected by copyright. All rights reserved.

Author Manuscript

JM-MLMM assumes following joint distribution when both outcome and covariates have missing data

$$\begin{aligned}\begin{pmatrix}\boldsymbol{x}_i\\ \boldsymbol{y}_i\end{pmatrix}|\boldsymbol{t}_i\Big) &= \begin{pmatrix}\beta_{0(x)}+\beta_{1(x)}\boldsymbol{t}_i+\boldsymbol{b}_{0(x)i}+\boldsymbol{b}_{1(x)i}\boldsymbol{t}_i+\boldsymbol{\varepsilon}_{1i}\\ \beta_{0(y)}+\beta_{1(y)}\boldsymbol{t}_i+\boldsymbol{b}_{0(y)i}+\boldsymbol{b}_{1(y)i}\boldsymbol{t}_i+\boldsymbol{\varepsilon}_{2i}\end{pmatrix}\\ &= \begin{pmatrix}\mu_{ix}\mathbf{1}\\ \mu_{iy}\mathbf{1}\end{pmatrix}+\begin{pmatrix}\boldsymbol{Z}_i\boldsymbol{b}_{(x)i}\\ \boldsymbol{Z}_i\boldsymbol{b}_{(y)i}\end{pmatrix}+\begin{pmatrix}\boldsymbol{\varepsilon}_{1i}\\ \boldsymbol{\varepsilon}_{2i}\end{pmatrix}\end{aligned}$$

where $\boldsymbol{Z}_i = (\mathbf{1}, \boldsymbol{t}_i)$ and $\begin{pmatrix}\boldsymbol{b}_{(x)i}\\ \boldsymbol{b}_{(y)i}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \boldsymbol{\psi} = \begin{pmatrix}\psi_{11} & \psi_{12}\\ \psi_{21} & \psi_{22}\end{pmatrix}\right)$ and $\begin{pmatrix}\boldsymbol{\varepsilon}_{1i}\\ \boldsymbol{\varepsilon}_{2i}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\right)$

Thus the joint distribution of ($\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ conditional on $\boldsymbol{t}_i$) can be written as (Resche-Rigon and White, 2016)

$$\begin{pmatrix}\boldsymbol{x}_i\\ \boldsymbol{y}_i\end{pmatrix}|\boldsymbol{t}_i\Big) \sim N\left(\begin{pmatrix}\mu_{ix}\mathbf{1}\\ \mu_{iy}\mathbf{1}\end{pmatrix}, \begin{pmatrix}\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{11}\boldsymbol{I} & \boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{12}\boldsymbol{I}\\ \boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{12}\boldsymbol{I} & \boldsymbol{Z}_i\psi_{22}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{22}\boldsymbol{I}\end{pmatrix}\right),$$

where $\mu_{ix} = \beta_{0(x)} + \beta_{1(x)}\boldsymbol{t}_i$ and $\mu_{iy} = \beta_{0(y)} + \beta_{1(y)}\boldsymbol{t}_i$.

Now the conditional expectation and conditional variance are

$$E[\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{t}_i] = \mu_{iy}\mathbf{1} + (\boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{12}\boldsymbol{I})(\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{11}\boldsymbol{I})^{-1}(\boldsymbol{x}_i-\mu_{ix}\mathbf{1}) \tag{17}$$

and

$$Var[\boldsymbol{y}_{1i}|\boldsymbol{x}_i,\boldsymbol{t}_i] = (\boldsymbol{Z}_i\psi_{22}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{22}\boldsymbol{I}) - (\boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{12}\boldsymbol{I})(\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{11}\boldsymbol{I})^{-1}(\boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{12}\boldsymbol{I}) \tag{18}$$

Now, for simplicity in algebra we omitted notation $\boldsymbol{I}$ associated with variance covariance matrices in the following equations

$$\begin{aligned}(\boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{12})(\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{11})^{-1} &= (\boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{12})(\Sigma_{11}^{-1}-\Sigma_{11}^{-1}\boldsymbol{Z}_i\psi_{11}(\boldsymbol{I}+\boldsymbol{Z}_i^{\mathrm{T}}\Sigma_{11}^{-1}\boldsymbol{Z}_i\psi_{11})^{-1}\boldsymbol{Z}_i^{\mathrm{T}}\Sigma_{11}^{-1})\\ &= (\boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{12})(\Sigma_{11}^{-1}-\Sigma_{11}^{-1}\boldsymbol{Z}_i(\psi_{11}^{-1}+\boldsymbol{Z}_i^{\mathrm{T}}\Sigma_{11}^{-1}\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i^{\mathrm{T}}\Sigma_{11}^{-1})\\ &= (\boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}+\Sigma_{12})(\Sigma_{11}^{-1}-\Sigma_{11}^{-1}\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}}(\Sigma_{11}+\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}})^{-1})\\ &= \boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}\Sigma_{11}^{-1}-\boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}\Sigma_{11}^{-1}\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}}(\Sigma_{11}+\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}})^{-1}+\\ &\quad \Sigma_{12}\Sigma_{11}^{-1}-\Sigma_{12}\Sigma_{11}^{-1}\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}}(\Sigma_{11}+\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}})^{-1}\\ &= (\boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}\Sigma_{11}^{-1}(\Sigma_{11}+\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}})-\boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}\Sigma_{11}^{-1}\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}}-\\ &\quad \Sigma_{12}\Sigma_{11}^{-1}\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}})(\Sigma_{11}+\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}})^{-1}+\Sigma_{12}\Sigma_{11}^{-1}\\ &= (\boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i^{\mathrm{T}}-\Sigma_{12}\Sigma_{11}^{-1}\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}})(\Sigma_{11}+\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}})^{-1}+\Sigma_{12}\Sigma_{11}^{-1}\\ &= (\boldsymbol{I}-\Sigma_{12}\Sigma_{11}^{-1})\boldsymbol{Z}_i\psi_{12}\boldsymbol{Z}_i(\Sigma_{11}+\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}})^{-1}+\Sigma_{12}\Sigma_{11}^{-1}\\ &= (\Sigma_{11}-\Sigma_{12})\Sigma_{11}^{-1}\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}}(\Sigma_{11}+\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}})^{-1}+\beta_x\boldsymbol{I}\end{aligned}$$

where $\beta_x = \Sigma_{12}\Sigma_{11}^{-1}$.

Thus the conditional mean is given as

$$\begin{aligned}E[\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{t}_i] &= \mu_y\mathbf{1}+((\Sigma_{11}-\Sigma_{12})\Sigma_{11}^{-1}\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}}(\Sigma_{11}+\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}})^{-1}+\beta_x\boldsymbol{I})(\boldsymbol{x}_i-\mu_x\mathbf{1})\\ &= (\mu_y-\beta_x\mu_x)\mathbf{1}+\beta_x\boldsymbol{x}_i+(\Sigma_{11}-\Sigma_{12})\Sigma_{11}^{-1}\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}}(\Sigma_{11}+\boldsymbol{Z}_i\psi_{11}\boldsymbol{Z}_i^{\mathrm{T}})^{-1}(\boldsymbol{x}_i-\mu_x\mathbf{1})\\ &= (\mu_y-\beta_x\mu_x)\mathbf{1}+\beta_x\boldsymbol{x}_i+(\Sigma_{11}-\Sigma_{12})\Sigma_{11}^{-1}\boldsymbol{Z}_i\boldsymbol{b}_{(x)i}\\ &= (\beta_{0(x)}+\beta_{1(x)}\boldsymbol{t}_i+\beta_x\boldsymbol{x}_i)\mathbf{1}+(\Sigma_{11}-\Sigma_{12})\Sigma_{11}^{-1}\boldsymbol{Z}_i\boldsymbol{b}_{(x)i}-\beta_x\mu_x,\end{aligned}$$

This article is protected by copyright. All rights reserved.

where $\boldsymbol{b}_{(x)i}$ is the estimates of random slope corresponding to the model $\boldsymbol{x}_i = \mu_{(ix)} + \boldsymbol{Z}_i\boldsymbol{b}_{(x)i} + \varepsilon_i$.

Thus the conditional expectation of $\boldsymbol{y}_i$ depends on $\boldsymbol{x}_i$ and random slopes of $\boldsymbol{t}_i$ on $\boldsymbol{x}_i$. Now can obtain the the substantive model if $(\Sigma_{22} - \Sigma_{12})\Sigma_{22}^{-1} = I$ or $\Sigma_{12} = 0$, as we have

$$E[\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{t}_i] \quad = \quad (\beta_{0(y)} + \beta_{1(y)}\boldsymbol{t}_i + \beta_x\boldsymbol{x}_i)\mathbf{1} + b_{0(x)i} + b_{1(x)i}\boldsymbol{t}_i - \beta_x\mu_x.$$

Thus the JM-MLMM impute missing values from a joint model which is more general than joint model implied by substantive model. Hence JM-MLMM is compatible with the analysis model.

Author Manuscript

This article is protected by copyright. All rights reserved.

A.3 Joint modelling: FJ

The JM-MLMM-LN approach assumes a multivariate linear random intercept model for the joint distribution of $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ where $\boldsymbol{z}$ is a latent normal variable for time $\boldsymbol{t}_i$

$$\begin{aligned}\begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{x}_i \\ \boldsymbol{t}_i \end{pmatrix} &= \begin{pmatrix} \beta_{0(y)} + \boldsymbol{b}_{(y)0i} + \boldsymbol{\varepsilon}_{(y)i} \\ \beta_{0(x)} + \boldsymbol{b}_{(x)0i} + \boldsymbol{\varepsilon}_{(x)i} \\ \beta_{0(t)} + \boldsymbol{b}_{(t)0i} + \boldsymbol{\varepsilon}_{(t)i} \end{pmatrix} \\ &= \begin{pmatrix} \mu_x \\ \mu_y \\ \mu_t \end{pmatrix} + \begin{pmatrix} \boldsymbol{b}_{(x)0i} \\ \boldsymbol{b}_{(y)0i} \\ \boldsymbol{b}_{(t)0i} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_{(x)i} \\ \boldsymbol{\varepsilon}_{(y)i} \\ \boldsymbol{\varepsilon}_{(t)i} \end{pmatrix}\end{aligned}$$

where $\begin{pmatrix} \boldsymbol{b}_{(y)0i} \\ \boldsymbol{b}_{(x)0i} \\ \boldsymbol{b}_{(t)0i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{21} & \psi_{22} & \psi_{23} \\ \psi_{31} & \psi_{32} & \psi_{33} \end{pmatrix}\right)$ and $\begin{pmatrix} \boldsymbol{\varepsilon}_{(y)i} \\ \boldsymbol{\varepsilon}_{(x)i} \\ \boldsymbol{\varepsilon}_{(t)i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{pmatrix}\right)$

Thus the joint distribution between ($\boldsymbol{y}_i$, $\boldsymbol{x}_i$ and $\boldsymbol{t}_i$) can be written as

$$\begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{x}_i \\ \boldsymbol{t}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1\boldsymbol{1} \\ \mu_y\boldsymbol{1} \\ \mu_t\boldsymbol{1} \end{pmatrix}, \begin{pmatrix} \psi_{11}\boldsymbol{J} + \Sigma_{11}\boldsymbol{I} & \psi_{12}\boldsymbol{J} + \Sigma_{12}\boldsymbol{I} & \psi_{13}\boldsymbol{J} + \Sigma_{13}\boldsymbol{I} \\ \psi_{12}\boldsymbol{J} + \Sigma_{12}\boldsymbol{I} & \psi_{22}\boldsymbol{J} + \Sigma_{22}\boldsymbol{I} & \psi_{23}\boldsymbol{J} + \Sigma_{23}\boldsymbol{I} \\ \psi_{13}\boldsymbol{J} + \Sigma_{13}\boldsymbol{I} & \psi_{23}\boldsymbol{J} + \Sigma_{23}\boldsymbol{I} & \psi_{33}\boldsymbol{J} + \Sigma_{33}\boldsymbol{I} \end{pmatrix}\right).$$

To ease in exposition let $\boldsymbol{w}_i = \begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{t}_i \end{pmatrix}$ and consequently the joint distribution between $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$ can be written as

$$\begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{w}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu^1 \otimes \boldsymbol{1} \\ \mu^2 \otimes \boldsymbol{1} \end{pmatrix}, \begin{pmatrix} \psi^{11} \otimes \boldsymbol{J} + \Sigma^{11} \otimes \boldsymbol{I} & \psi^{12} \otimes \boldsymbol{J} + \Sigma^{12} \otimes \boldsymbol{I} \\ \psi^{21} \otimes \boldsymbol{J} + \Sigma^{21} \otimes \boldsymbol{I} & \psi^{22} \otimes \boldsymbol{J} + \Sigma^{22} \otimes \boldsymbol{I} \end{pmatrix}\right),$$

where $\boldsymbol{1}$ is n_i-vector of ones, $\boldsymbol{J}$ is a $n_i \times n_i$ matrix of ones, and $\boldsymbol{I}$ is the identity matrix of size n_i and $\mu^1 = \mu_y$, $\mu^2 = \begin{pmatrix} \mu_x \\ \mu_z \end{pmatrix}$, $\psi^{11} = \psi_{11}$, $\Sigma^{11} = \Sigma_{11}$,$\psi^{21} = (\psi^{12})^{\mathrm{T}} = \begin{pmatrix} \psi_{21} \\ \psi_{31} \end{pmatrix}$,$\Sigma^{21} = (\Sigma^{12})^{\mathrm{T}} = \begin{pmatrix} \Sigma_{21} \\ \Sigma_{31} \end{pmatrix}$ and

$\psi^{22} = \begin{pmatrix} \psi_{22} & \psi_{23} \\ \psi_{32} & \psi_{33} \end{pmatrix}$, $\Sigma^{22} = \begin{pmatrix} \Sigma_{22} & \Sigma_{23} \\ \Sigma_{32} & \Sigma_{33} \end{pmatrix}$.

Now the conditional expectation $\boldsymbol{y}_i$ given $\boldsymbol{w}_i$ is

$$E[\boldsymbol{y}_i|\boldsymbol{w}_i] = \mu^1 \otimes \boldsymbol{1} + \left(\ \psi^{12} \otimes \boldsymbol{J} + \Sigma^{12} \otimes \boldsymbol{I}\ \right)\left(\ \psi^{22} \otimes \boldsymbol{J} + \Sigma^{22} \otimes \boldsymbol{I}\ \right)^{-1}\left(\ \boldsymbol{v}_i - \mu^2 \otimes \boldsymbol{1}\ \right) \quad (19)$$

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

Now

$$\begin{aligned}
\left(\ \psi^{12}\otimes\boldsymbol{J}+\Sigma^{12}\otimes\boldsymbol{I}\ \right)\left(\ \psi^{22}\otimes\boldsymbol{J}+\Sigma^{22}\otimes\boldsymbol{I}\ \right)^{-1} &= \left(\ \psi^{12}\otimes\boldsymbol{J}+\Sigma^{12}\otimes\boldsymbol{I}\ \right)\left(-(n\psi^{22}+\Sigma^{22})^{-1}\psi^{22}(\Sigma^{22})^{-1}\right)\otimes\boldsymbol{J} \\
&\quad +(\Sigma^{22})^{-1}\otimes\boldsymbol{I} \\
&= -\psi^{12}(n\psi^{22}+\Sigma^{22})^{-1}\psi^{22}(\Sigma^{22})^{-1}\otimes n\boldsymbol{J}+\Sigma^{12}(\Sigma^{22})^{-1}\otimes\boldsymbol{I} \\
&\quad -\Sigma^{12}(n\psi^{22}+\Sigma^{22})^{-1}\psi^{22}(\Sigma^{22})^{-1}\otimes\boldsymbol{J}+\psi^{12}(\Sigma^{22})^{-1}\otimes\boldsymbol{J} \\
&= (-n\psi^{12}(n\psi^{22}+\Sigma^{22})^{-1}\psi^{22}-\Sigma^{12}(n\psi^{22}+\Sigma^{22})^{-1}\psi^{22} \\
&\quad +\psi^{12})(\Sigma^{22})^{-1}\otimes\boldsymbol{J}+\Sigma^{12}(\Sigma^{22})^{-1}\otimes\boldsymbol{I} \\
&= (-n\psi^{12}-\Sigma^{12}+\psi^{12}(\psi^{22})^{-1}(n\psi^{22}+\Sigma^{22})) \\
&\quad (n\psi^{22}+\Sigma^{22})^{-1}\psi^{22}(\Sigma^{22})^{-1}\otimes\boldsymbol{J}+\Sigma^{12}(\Sigma^{22})^{-1}\otimes\boldsymbol{I} \\
&= (-n\psi^{12}-\Sigma^{12}+n\psi^{12}(\psi^{22})^{-1}\psi^{22}+\psi^{12}(\psi^{22})^{-1}\Sigma^{22}) \\
&\quad (n\psi^{22}+\Sigma^{22})^{-1}\psi^{22}(\Sigma^{22})^{-1}\otimes\boldsymbol{J}+\Sigma^{12}(\Sigma^{22})^{-1}\otimes\boldsymbol{I} \\
&= (\psi^{12}(\psi^{22})^{-1}\Sigma^{22})-\Sigma^{12})(n\psi^{22}+\Sigma^{22})^{-1}\psi^{22}(\Sigma^{22})^{-1}\otimes\boldsymbol{J} \\
&\quad +\Sigma^{12}(\Sigma^{22})^{-1}\otimes\boldsymbol{I} \\
&= (\psi^{12}\Sigma^{22}-\Sigma^{12}\psi^{22})(n\psi^{22}+\Sigma^{22})^{-1}(\Sigma^{22})^{-1}\otimes\boldsymbol{J} \\
&\quad +\Sigma^{12}(\Sigma^{22})^{-1}\otimes\boldsymbol{I} \\
&= \boldsymbol{\alpha}(n_i)\otimes\boldsymbol{J}+\beta\otimes\boldsymbol{I},
\end{aligned}$$

where $\beta=(\beta_y,\beta_t)=\Sigma^{12}(\Sigma^{22})^{-1}$ and $\boldsymbol{\alpha}(n_i)=(\psi^{12}\Sigma^{22}-\Sigma^{12}\psi^{22})(n\psi^{22}+\Sigma^{22})^{-1}(\Sigma^{22})^{-1}$.

Hence,

$$E[\boldsymbol{y}_i|\boldsymbol{w}_i] = \mu^1\otimes\mathbf{1}+(\boldsymbol{\alpha}(n_i)\otimes\boldsymbol{J}+\beta\otimes\boldsymbol{I})\left(\boldsymbol{w}_i-\mu^2\otimes\mathbf{1}\right)$$

Now considering the identity and from the fact that $vec(w_i)=\boldsymbol{w}_i$, we can show that

$$\begin{aligned}
E[\boldsymbol{y}_i|\boldsymbol{w}_i] &= \mu^1\otimes\mathbf{1}+vec(\boldsymbol{J}(w_i-\mu^2\otimes\mathbf{1})\boldsymbol{\alpha}(n_i)^{\mathrm{T}})+(w_i-\mu^2\otimes\mathbf{1})\beta^{\mathrm{T}}) \\
&= \mu_y\mathbf{1}+n_i(\bar{w}_i-\mu^2)\otimes\mathbf{1}\boldsymbol{\alpha}(n_i)^{\mathrm{T}}+(w_i-\mu^2\otimes\mathbf{1})\beta^{\mathrm{T}}
\end{aligned}$$

Thus we have

$$E[\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{t}_i] = \beta_{0(y)}+n_i(\bar{\boldsymbol{x}}_i-\mu_x)\boldsymbol{\alpha}(n_i)^{\mathrm{T}}+n_i(\bar{\boldsymbol{t}}_i-\mu_t)\boldsymbol{\alpha}(n_i)^{\mathrm{T}}+(\boldsymbol{x}_i-\mu_x)\beta_x+(\boldsymbol{t}_i-\mu_t)\beta_t$$

And the conditional variance is

$$\begin{aligned}
var[\boldsymbol{y}_i|\boldsymbol{w}_i] &= \left(\ \psi^{11}\otimes\boldsymbol{J}+\Sigma^{11}\otimes\boldsymbol{I}\ \right)-\left(\ \psi^{12}\otimes\boldsymbol{J}+\Sigma^{12}\otimes\boldsymbol{I}\ \right)\left(\ \psi^{22}\otimes\boldsymbol{J}+\Sigma^{22}\otimes\boldsymbol{I}\ \right)^{-1}\left(\ \psi^{12}\otimes\boldsymbol{J}+\Sigma^{12}\otimes\boldsymbol{I}\ \right) \\
&= \eta(n_i)\otimes\boldsymbol{J}+\delta\otimes\boldsymbol{I}
\end{aligned}$$

where $\eta(n_i)=\psi^{22}-\psi^{12}(\Sigma^{11})^{-1}\Sigma^{12}+(n_i\psi^{21}+\Sigma^{21})(n_i\psi^{11}+\Sigma^{11})^{-1}(\psi^{11}(\Sigma^{11})^{-1}\Sigma^{12}-\psi^{12})$ and $\delta=\Sigma^{22}-\Sigma^{12}(\Sigma^{11})^{-1}\Sigma^{12}$

As the joint distribution implied by JM-MLMM-LN doesn't contain the conditional substantive model (1) as the corresponding conditional model, hence it would not be compatible.

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This article is protected by copyright. All rights reserved.

Author Manuscript

A.4 JM-MLMM for cluster-correlated data

Suppose we are interested in substantive model (2), with slight abuse of notations we can write the substantive model as

$$\boldsymbol{y}_i|\boldsymbol{x}_{1i},\boldsymbol{x}_{2i} = \alpha_{0i} + \alpha_1\boldsymbol{x}_{1i} + \alpha_{2i}\boldsymbol{x}_{2i} + \xi_i, \tag{20}$$

where $\alpha_{0i} = \alpha_0 + \boldsymbol{a}_{0i}$, $\alpha_{2i} = \alpha_2 + \boldsymbol{a}_{2i}$ and $\xi_i \sim N(0, \Sigma_{iy} = \sigma^2_\xi \boldsymbol{I})$. Now suppose that the covariate $\boldsymbol{x}_2$ follows the following LMM:

$$\boldsymbol{x}_{2i}|x_{1i} = \delta_{0i} + \delta_{1i}\boldsymbol{x}_{1i} + \eta_i, \tag{21}$$

where $\eta_i \sim N(0, \Sigma_{ix_2} = \sigma^2_{\eta_i}\boldsymbol{I})$ As both of the $(\boldsymbol{y}_i|\boldsymbol{x}_{1i},\boldsymbol{x}_{2i})$ and $(\boldsymbol{x}_{2i}|x_{1i})$ are Gaussian, their joint distribution will also be Gaussian and can be written as

$$\begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{x}_{2i} \end{pmatrix} |\boldsymbol{x}_{1i}\Bigg) = N\left(\begin{pmatrix} \alpha_{0i} + \alpha_1\boldsymbol{x}_{1i} + \alpha_{2i}(\delta_{0i} + \delta_{1i}\boldsymbol{x}_{1i}) \\ \delta_{0i} + \delta_{1i}\boldsymbol{x}_{1i} \end{pmatrix}, \begin{pmatrix} \alpha_{2i}\Sigma_{ix}\alpha^{\mathrm{T}}_{2i} + \Sigma_{iy} & \alpha_{2i}\Sigma_{ix} \\ \Sigma_{ix_2}\alpha_{2i} & \Sigma_{ix_2} \end{pmatrix}\right), \tag{22}$$

Now the conditional distribution of $(\boldsymbol{x}_{2i}|\boldsymbol{x}_{1i},\boldsymbol{y}_i)$ can be written as

$$(\boldsymbol{x}_{2i}|\boldsymbol{x}_{1i},\boldsymbol{y}_i) = N\left(\Sigma_{ix_2|y}\left[\alpha^{\mathrm{T}}_{2i}\Sigma^{-1}_{iy}(\boldsymbol{y}_i - (\alpha_{0i} + \alpha_1\boldsymbol{x}_{1i})) + \Sigma^{-1}_{ix_2}(\delta_{0i} + \delta_{1i}\boldsymbol{x}_{1i})\right], \Sigma_{ix_2|y} = (\alpha^{\mathrm{T}}_{2i}\Sigma^{-1}_{iy}\alpha_{2i} + \Sigma^{-1}_{ix_2})^{-1}\right) \tag{23}$$

When only covariates $\boldsymbol{x}_2$ contains missing data, the imputation model followed by JM-MLMM is given as

$$\boldsymbol{x}_{2i}|\boldsymbol{x}_{1i},\boldsymbol{y}_i = \alpha_{0i(x_2)} + \alpha_{1(x_2)}\boldsymbol{x}_{1i} + \alpha_{2i(x_2)}\boldsymbol{y}_i + \xi_{(x_2)i}, \tag{24}$$

This imputation model will be compatible to the substantive model if $\Sigma^{-1}_{ix_2|y}\alpha_{2i(x_2)} = \Sigma_{ix_2|y}\alpha^{\mathrm{T}}_{2i}\Sigma^{-1}_{iy}$

This article is protected by copyright. All rights reserved.

Author Manuscript

B Additional Results

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This article is protected by copyright. All rights reserved.

Table B1 Simulation results for the analysis of longitudinal data using simulation scenario (i), i.e., scenario with data missing in the covariate only with 1000 samples observed for 5 time periods

Regression Parameters	True value	Available data	JM-MVN	JM-MLMM	JM-FJ	JM-FJ-het	JM-SMC	JM-SMC-het	FCS-standard	FCS-LMM	FCS-LMM-het
cbmi, $\widehat{\beta_1}$	-0.020	-0.184	-0.197	-0.201	-0.196	-0.186	-0.196	-0.199	-0.197	-0.201	-0.202
$\widehat{\beta_1}$ rbias (%)		0.079	0.015	0.006	0.019	0.069	0.017	0.002	0.013	0.006	0.011
Model SE		0.018	0.020	0.020	0.019	0.020	0.019	0.019	0.020	0.019	0.020
Empirical SE		0.018	0.020	0.020	0.019	0.019	0.019	0.019	0.020	0.020	0.020
95% Coverage		0.849	0.950	0.945	0.944	0.906	0.939	0.948	0.949	0.943	0.947
cage, $\widehat{\beta_2}$	-0.100	-0.090	-0.100	-0.100	-0.100	-0.101	-0.101	-0.100	-0.100	-0.100	-0.100
$\widehat{\beta_1}$ rbias (%)		0.098	0.003	0.001	0.005	0.016	0.007	0.004	0.003	0.001	0.001
Model SE		0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
Empirical SE		0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
95% Coverage		0.449	0.952	0.953	0.950	0.931	0.952	0.952	0.951	0.957	0.954
Variance components											
G_{00} est	0.360	0.337	0.360	0.367	0.362	0.363	0.362	0.361	0.360	0.367	0.365
G_{00}rbias (%)		0.064	0.001	0.019	0.006	0.009	0.006	0.002	0.001	0.019	0.015
G_{01} est	-0.012	-0.012	-0.012	-0.013	-0.012	-0.012	-0.012	-0.012	-0.012	-0.013	-0.012
G_{01}rbias (%)		0.010	0.005	0.064	0.007	0.025	0.033	0.021	0.006	0.061	0.045
G_{11} est	0.004	0.003	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.040
G_{11}rbias (%)		0.143	0.001	0.028	0.002	0.006	0.016	0.011	0.002	0.027	0.018
Residual error, $\widehat{\sigma}^2_{\varepsilon}$	0.436	0.414	0.435	0.434	0.435	0.436	0.435	0.434	0.435	0.434	0.434
$\widehat{\sigma}_{\varepsilon}$ rbias (%)		0.050	0.001	0.004	<0.001	0.001	0.001	0.004	0.001	0.004	0.004

This article is protected by copyright. All rights reserved.

Author Manuscript

Author Manuscript

Table B2 Simulation results for the analysis of longitudinal data using simulation scenario (ii), i.e., scenario with data missing in both covariate and outcome with 1000 samples observed for 5 time periods

Regression Parameters	True value	Available data	JM-MVN	JM-MLMM	JM-FJ	JM-FJ-het	JM-SMC	JM-SMC-het	FCS-standard	FCS-LMM	FCS-LMM-het
cbmi, $\widehat{\beta_1}$	-0.200	-0.184	-0.196	-0.195	-0.197	-0.184	-0.196	-0.199	-0.196	-0.200	-0.201
$\widehat{\beta_1}$ rbias (%)		0.082	0.019	0.026	0.012	0.081	0.018	0.005	0.019	0.001	0.004
Model SE		0.020	0.023	0.023	0.023	0.024	0.022	0.022	0.023	0.022	0.022
Empirical SE		0.020	0.023	0.023	0.023	0.021	0.022	0.022	0.023	0.022	0.022
95% Coverage		0.861	0.944	0.938	0.946	0.919	0.945	0.949	0.948	0.950	0.951
cage, $\widehat{\beta_2}$	-0.100	-0.089	-0.101	-0.101	-0.101	-0.102	-0.101	-0.101	-0.101	-0.100	-0.100
$\widehat{\beta_1}$ rbias (%)		-0.111	0.006	0.009	0.007	0.019	0.009	0.007	0.006	0.003	0.001
Model SE		0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
Empirical SE		0.006	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
95% Coverage		0.501	0.953	0.956	0.941	0.953	0.948	0.951	0.954	0.963	0.948
Variance components											
G_{00} est	0.360	0.333	0.355	0.397	0.345	0.426	0.343	0.341	0.356	0.406	0.400
G_{00}rbias (%)		0.074	0.013	0.102	0.042	0.154	0.046	0.052	0.012	0.128	0.112
G_{01} est	-0.012	-0.011	-0.011	-0.019	-0.003	-0.014	-0.010	-0.010	-0.011	-0.020	-0.016
G_{01} rbias (%)		0.034	0.078	0.629	0.729	0.125	0.138	0.155	0.076	0.689	0.309
G_{11} est	0.004	0.003	0.004	0.005	0.002	0.003	0.004	0.004	0.004	0.005	0.004
G_{11} rbias(%)		0.162	0.006	0.372	0.465	0.162	0.035	0.041	0.004	0.385	0.109
Residual error, $\widehat{\sigma}^2_\varepsilon$	0.436	0.414	0.442	0.429	0.454	0.483	0.437	0.436	0.442	0.426	0.434
$\widehat{\sigma}_\varepsilon$ rbias (%)		0.049	0.015	0.016	0.041	0.110	0.003	<0.001	0.014	0.022	0.004

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This article is protected by copyright. All rights reserved.

Table B3 Simulation results for the analysis of clustered data using simulation scenario (iii),i.e., scenario with data missing in the covariate only with small samples (100 clusters with 2 to 10 observations per cluster)

Regression Parameters	True value	Available data	JM-MLMM	JM-FJ	JM-FJ-het	JM-SMC	JM-SMC-het	FCS LMM	FCS-LMM-het
SEIFA, $\widehat{\gamma}_1$	0.250	0.191	0.246	0.250	0.191	0.250	0.250	0.246	0.248
$\widehat{\gamma}_2$ rbias (%)		0.235	0.016	0.001	0.235	<0.001	<0.001	0.015	0.008
Model SE		0.059	0.058	0.058	0.059	0.058	0.058	0.058	0.058
Empirical SE		0.061	0.059	0.060	0.061	0.060	0.060	0.059	0.060
95% Coverage		0.816	0.946	0.939	0.822	0.940	0.938	0.946	0.945
cbmi, $\widehat{\gamma}_2$	-0.200	-0.177	-0.193	-0.202	-0.177	-0.196	-0.195	-0.192	-0.193
$\widehat{\gamma}_1$ rbias (%)		0.115	0.034	0.012	0.115	0.021	0.021	0.042	0.036
Model SE		0.049	0.057	0.053	0.050	0.056	0.056	0.057	0.056
Empirical SE		0.051	0.054	0.057	0.051	0.058	0.058	0.054	0.054
95% Coverage		0.908	0.961	0.927	0.911	0.936	0.943	0.958	0.955
Variance components									
G_{00} est	0.400	0.331	0.388	0.402	0.340	0.392	0.392	0.388	0.390
G_{00}rbias (%)		0.173	0.030	0.005	0.149	0.018	0.018	0.030	0.025
G_{11} est	0.200	0.154	0.225	0.136	0.160	0.195	0.195	0.226	0.216
G_{11} rbias(%)		0.229	0.127	0.322	0.198	0.023	0.022	0.131	0.080
Residual error, $\widehat{\sigma}_e$	0.900	0.846	0.890	0.908	0.846	0.897	0.898	0.890	0.892
$\widehat{\sigma}_e$ rbias (%)		0.060	0.011	0.008	0.059	0.003	0.003	0.011	0.008

This article is protected by copyright. All rights reserved.

Author Manuscript

Author Manuscript

Table B4 Simulation results for the analysis of clustered data using simulation scenario (iv), i.e., senario with data missing in covariate associated with random slope and outcome with small samples (100 clusters with 2 to 10 observations per cluster)

Regression Parameters	True value	Available data	JM-MLMM	JM-FJ	JM-FJ-het	JM-SMC	JM-SMC-het	FCS-LMM	FCS-LMM-het
SEIFA, $\widehat{\gamma}_1$	0.250	0.191	0.250	0.249	0.191	0.250	0.251	0.247	0.250
$\widehat{\gamma}_2$ rbias (%)		0.234	0.002	0.002	0.234	0.001	0.002	0.010	0.001
Model SE		0.067	0.066	0.066	0.068	0.065	0.065	0.067	0.070
Empirical SE		0.067	0.066	0.065	0.067	0.065	0.065	0.065	0.066
95% Coverage		0.843	0.959	0.956	0.852	0.958	0.956	0.960	0.961
cbmi, $\widehat{\gamma}_2$	-0.200	-0.181	-0.205	-0.203	-0.181	-0.194	-0.194	-0.188	-0.190
$\widehat{\gamma}_1$ rbias (%)		0.094	0.027	0.016	0.094	0.030	0.029	0.058	0.048
Model SE		0.059	0.063	0.062	0.059	0.068	0.068	0.070	0.070
Empirical SE		0.060	0.065	0.065	0.060	0.065	0.065	0.062	0.062
95% Coverage		0.934	0.935	0.942	0.934	0.958	0.961	0.968	0.974
Variance components									
G_{00} est	0.400	0.317	0.416	0.408	0.330	0.383	0.384	0.396	0.420
G_{00}rbias (%)		0.208	0.039	0.020	0.174	0.042	0.041	0.009	0.050
G_{11} est	0.200	0.148	0.112	0.112	0.154	0.219	0.219	0.273	0.254
G_{11} rbias(%)		0.262	0.439	0.441	0.225	0.096	0.095	0.366	0.274
Residual error, $\widehat{\sigma}_e$	0.900	0.844	0.909	0.912	0.845	0.896	0.896	0.884	0.898
$\widehat{\sigma}_e$ rbias (%)		0.062	0.010	0.013	0.061	0.004	0.004	0.017	0.002

© 2010 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

This article is protected by copyright. All rights reserved.

University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Huque, MH;Moreno-Betancur, M;Quartagno, M;Simpson, JA;Carlin, JB;Lee, KJ

Title:
Multiple imputation methods for handling incomplete longitudinal and clustered data where the target analysis is a linear mixed effects model

Date:
2020-01-09

Citation:
Huque, M. H., Moreno-Betancur, M., Quartagno, M., Simpson, J. A., Carlin, J. B. & Lee, K. J. (2020). Multiple imputation methods for handling incomplete longitudinal and clustered data where the target analysis is a linear mixed effects model. BIOMETRICAL JOURNAL, 62 (2), pp.444-466. https://doi.org/10.1002/bimj.201900051.

Persistent Link:
http://hdl.handle.net/11343/275250